CLC **Gene** Workbench

User manual

User manual for

CLC Gene Workbench 2.0

Windows, Mac OS X and Linux

July 6, 2006

CLC bio
Gustav Wieds Vej 10
Dk-8000 Aarhus C
Denmark

Contents

I Introduction **8**

1 Introduction to *CLC Gene Workbench* **9**

1.1 Contact information 11
1.2 Download and installation 11
1.3 System requirements 14
1.4 Licenses 15
1.5 About CLC Workbenches 18
1.6 When the program is installed: Getting started 20
1.7 Network configuration 22
1.8 Adjusting the maximum amount of memory 22
1.9 The format of the user manual 24

2 Tutorials **25**

2.1 Tutorial: Starting up the program 26
2.2 Tutorial: View sequence 29
2.3 Tutorial: GenBank search and download 31
2.4 Tutorial: Align protein sequences 32
2.5 Tutorial: Create and modify a phylogenetic tree 34
2.6 Tutorial: Detect restriction sites 35
2.7 Tutorial: Sequence information 37
2.8 Tutorial: BLAST search 38
2.9 Tutorial: Primer design 41
2.10 Tutorial: Assembly 45
2.11 Tips and tricks for the experienced user 47

II Basic Program Functionalities **56**

3 User Interface **57**

3.1 Navigation Area . . . 58
3.2 View Area . . . 64
3.3 Zoom and selection in View Area . . . 69
3.4 Toolbox and Status Bar . . . 72
3.5 Workspace . . . 73
3.6 List of shortcuts . . . 74

4 User preferences **76**

4.1 General preferences . . . 77
4.2 Default View preferences . . . 77
4.3 Advanced preferences . . . 78
4.4 Export/import of preferences . . . 78
4.5 View preference style sheet . . . 78

5 Printing **82**

5.1 Selecting which part of the view to print . . . 82
5.2 Page setup . . . 83
5.3 Print preview . . . 83

6 Import/export of data and graphics **85**

6.1 Bioinformatic data formats . . . 85
6.2 External files . . . 90
6.3 Export graphics to files . . . 91
6.4 Copy/paste view output . . . 93

7 History **95**

7.1 Element history . . . 95

8 Handling of results **97**

8.1 How to handle results of analyses . . . 97

III Bioinformatics 100

9 Database search 101

9.1 GenBank search 101

9.2 Sequence web info 104

10 BLAST Search 107

10.1 BLAST Against NCBI Database 107

10.2 BLAST Against Local Database 113

10.3 Create Local BLAST Database 114

11 Viewing and editing sequences 117

11.1 View sequence 117

11.2 Sequence information 127

11.3 View as text 128

11.4 Creating a new sequence 129

11.5 Sequence Lists 130

11.6 Circular DNA 132

12 General sequence analyses 136

12.1 Dot plots 136

12.2 Shuffle sequence 146

12.3 Local complexity plot 147

12.4 Sequence statistics 149

12.5 Join sequences 156

12.6 Motif Search 157

12.7 Pattern Discovery 160

13 Nucleotide analyses 163

13.1 Convert DNA to RNA 163

13.2 Convert RNA to DNA 164

13.3 Reverse complements of sequences 165

13.4 Translation of DNA or RNA to protein 166

13.5 Annotate with SNPs 167

13.6 Find open reading frames 168

14 Protein analyses **171**

14.1 Protein charge 171
14.2 Hydrophobicity 174
14.3 Reverse translation from protein into DNA 179

15 Primers **184**

15.1 Primer design - an introduction 185
15.2 Setting parameters for primers and probes 187
15.3 Graphical display of primer information 189
15.4 Output from primer design 191
15.5 Standard PCR 192
15.6 Nested PCR 195
15.7 TaqMan 197
15.8 Sequencing primers 199
15.9 Alignment-based primer and probe design 200
15.10 Analyze primer properties 204
15.11 Match primer with sequence 205
15.12 Order primers 207

16 Assembly **208**

16.1 Importing and viewing trace data 208
16.2 Trim sequences 209
16.3 Assemble sequences 212
16.4 Assemble to reference sequence 213
16.5 Assemble to an existing contig 216
16.6 View and edit contigs 216

17 Cloning and cutting **221**

17.1 Molecular cloning - an introduction 221
17.2 Graphical display of *in silico* cloning 222
17.3 Restriction site analysis 230
17.4 Restriction enzyme lists 234

17.5 Gel electrophoresis 235

18 Sequence alignment **240**

18.1 Create an alignment 241

18.2 View alignments 246

18.3 Edit alignments 249

18.4 Join alignments 251

18.5 Bioinformatics explained: Multiple alignments 253

19 Phylogenetic trees **256**

19.1 Inferring phylogenetic trees 256

19.2 Bioinformatics explained: phylogenetics 259

IV Appendix **264**

A Comparison of workbenches **265**

B BLAST databases **268**

B.1 Peptide sequence databases 268

B.2 Nucleotide sequence databases 268

C Formats for import and export **270**

C.1 List of bioinformatic data formats 270

C.2 List of graphics data formats 271

Bibliography **272**

V Index **275**

Part I

Introduction

Chapter 1

Introduction to *CLC Gene Workbench*

Contents

1.1 Contact information . . . 11
1.2 Download and installation . . . 11
1.2.1 Program download . . . 11
1.2.2 Installation on Microsoft Windows . . . 12
1.2.3 Installation on Mac OS X . . . 13
1.2.4 Installation on Linux with an installer . . . 13
1.2.5 Installation on Linux with an RPM-package . . . 14
1.3 System requirements . . . 14
1.4 Licenses . . . 15
1.4.1 Demo license description . . . 15
1.4.2 Getting and activating the demo license . . . 15
1.4.3 Commercial license . . . 17
1.4.4 Upgrading from a demo license to a commercial license . . . 18
1.5 About CLC Workbenches . . . 18
1.5.1 New program feature request . . . 19
1.5.2 Report program errors . . . 19
1.5.3 Free vs. commercial workbenches . . . 19
1.6 When the program is installed: Getting started . . . 20
1.6.1 Basic concepts of using CLC Workbenches . . . 20
1.6.2 Quick start . . . 21
1.6.3 Import of example data . . . 22
1.7 Network configuration . . . 22
1.8 Adjusting the maximum amount of memory . . . 22
1.8.1 Microsoft Windows . . . 23
1.8.2 Mac OS X . . . 23
1.8.3 Linux . . . 23
1.9 The format of the user manual . . . 24
1.9.1 Text formats . . . 24

Welcome to *CLC Gene Workbench 2.0* — a software package supporting your daily bioinformatics work.

We strongly encourage you to read this user manual in order to get the best possible basis for working with the software package.

1.1 Contact information

The *CLC Gene Workbench 2.0* is developed by:

CLC bio A/S
Science Park Aarhus
Gustav Wieds Vej 10
8000 Aarhus C
Denmark

http://www.clcbio.com

VAT no.: DK 28 30 50 87

Telephone: +45 70 22 32 44
Fax: +45 86 20 12 22

E-mail: info@clcbio.com

If you have questions or comments regarding the program, you are welcome to contact our support function:

E-mail: support@clcbio.com

1.2 Download and installation

The *CLC Gene Workbench* is developed for Windows, Mac OS X and Linux. The software for either platform can be downloaded from http://www.clcbio.com/download.

Furthermore the program can be sent on a CD-Rom by regular mail. To receive the program by regular mail, please write an e-mail to support@clcbio.com, including your postal address.

1.2.1 Program download

The program is available for download on http://www.clcbio.com/download.

Before you download the program you are asked to fill in the **Download** dialog.

In the dialog you must choose:

- Which operating system you use
- Whether you want to include Java or not
  (this is necessary if you haven't already installed Java)
- Whether you would like to receive information about future releases

Depending on your operating system and your Internet browser, you are taken through some download options.

When the download of the installer (an application which facilitates the installation of the program) is complete, follow the platform specific instructions below to complete the installation procedure.

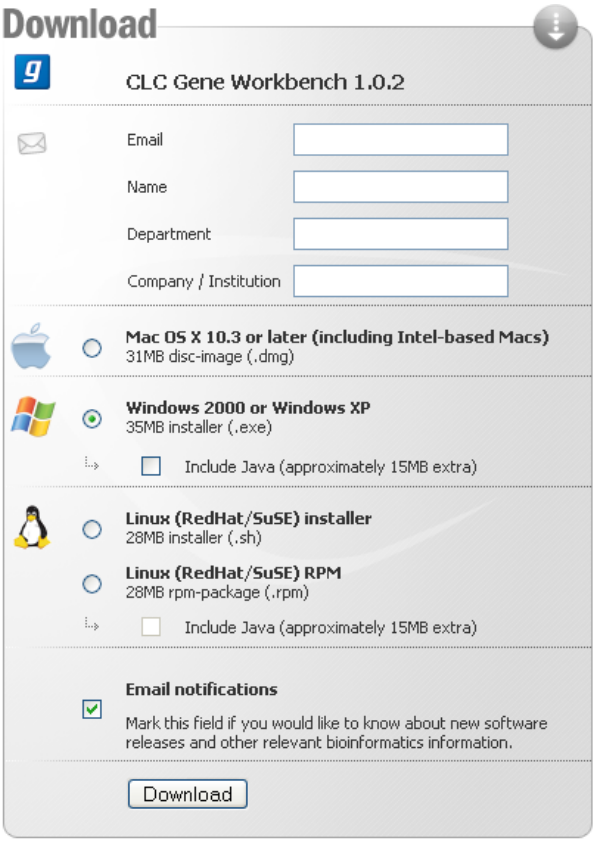


Figure 1.1: *Download dialog.*

1.2.2 Installation on Microsoft Windows

Starting the installation process is done in one of the following ways:

If you have downloaded an installer:
Locate the downloaded installer and double-click the icon.
The default location for downloaded files is your desktop.

If you are installing from a CD:
Insert the CD into your CD-ROM drive.

Choose the "Install *CLC Gene Workbench*" from the menu displayed.

If you already have Java installed on your computer you can choose "Install *CLC Gene Workbench* without Java".

Installing the program is done in the following steps:
(you must be connected to the Internet throughout the installation process.)

- On the welcome screen, click Next.
- Read and accept the License agreement and click Next.
- Choose where you would like to install the application and click Next.
- Choose a name for the Start Menu folder used to launch *CLC Gene Workbench* and click Next.

- Choose where you would like to create shortcuts for launching *CLC Gene Workbench* and click Next.
- Wait for the installation process to complete, choose whether you would like to launch *CLC Gene Workbench* right away, and click Finish.

When the installation is complete the program can be launched from the Start Menu or from one of the shortcuts you choose to create.

1.2.3 Installation on Mac OS X

Starting the installation process is done in one of the following ways:

If you have downloaded an installer:
Locate the downloaded installer and double-click the icon.
The default location for downloaded files is your desktop.

If you are installing from a CD:
Insert the CD into your CD-ROM drive and open it by double-clicking on the CD icon on your desktop.

Launch the installer by double-clicking on the "*CLC Gene Workbench*" icon.

Installing the program is done in the following steps:

- On the welcome screen, click Next.
- Read and accept the License agreement and click Next.
- Choose where you would like to install the application and click Next.
- Choose whether you would like to create desktop icon for launching *CLC Gene Workbench* and click Next.
- Wait for the installation process to complete, choose whether you would like to launch *CLC Gene Workbench* right away, and click Finish.

When the installation is complete the program can be launched from your Applications folder, or from the desktop shortcut you choose to create. If you like, you can drag the application icon to the dock for easy access.

1.2.4 Installation on Linux with an installer

Navigate to the directory containing the installer and execute it. This can be done by running a command similar to:

```
# sh CLCGeneWorkbench_1_0_2_JRE.sh.sh
```

If you are installing from a CD the installers are located in the "linux" directory.

Installing the program is done in the following steps:

- On the welcome screen, click Next.
- Read and accept the License agreement and click Next.
- Choose where you would like to install the application and click Next.
  For a system-wide installation you can choose for example /opt or /usr/local. If you do not have root privileges you can choose to install in your home directory.
- Choose where you would like to create symbolic links to the program
  DO NOT create symbolic links in the same location as the application.
  Symbolic links should be installed in a location which is included in your environment PATH. For a system-wide installation you can choose for example /usr/local/bin. If you do not have root privileges you can create a 'bin' directory in your home directory and install symbolic links there. You can also choose not to create symbolic links.
- Wait for the installation process to complete and click Finish.

If you choose to create symbolic links in a location which is included in your PATH, the program can be executed by running the command:

```
# clcgenewb
```

Otherwise you start the application by navigating to the location where you choose to install it and running the command:

```
# ./clcgenewb
```

1.2.5 Installation on Linux with an RPM-package

Navigate to the directory containing the rpm-package and install it using the rpm-tool by running a command similar to:

```
# rpm -ivh CLCGeneWorkbench_1_0_2_JRE.sh.rpm
```

If you are installing from a CD the rpm-packages are located in the "RPMS" directory.
Installation of RPM-packages usually requires root-privileges.

When the installation process is finished the program can be executed by running the command:

```
# clcgenewb
```

1.3 System requirements

The system requirements of *CLC Gene Workbench 2.0* are these:

- Windows 2000 or Windows XP
- Mac OS X 10.3 or newer
- Linux: Redhat or SuSE

- 256 MB RAM required
- 512 MB RAM recommended
- 1024 x 768 display recommended

1.4 Licenses

The license system of *CLC Gene Workbench 2.0* is based on a license key which is unique for the computer rather than for the user of the workbench.

1.4.1 Demo license description

We offer a fully functional demo version of *CLC Gene Workbench 2.0* to all users, free of charge.

Each user is entitled to four weeks demo of *CLC Gene Workbench 2.0*. In order to make your demo time as valuable as possible, the four weeks can be separated. You can e.g. try two weeks of the demo in January, and the next two weeks in March.

To prevent unauthorized use of the program, you must be connected to the Internet while starting up a demo version of *CLC Gene Workbench*. An additional online check will be conducted 24 hours after the launch of the workbench. After running *CLC Gene Workbench 2.0* for 24 hours, if you are not connected to the Internet, you will be met with the dialog shown in figure 1.2.

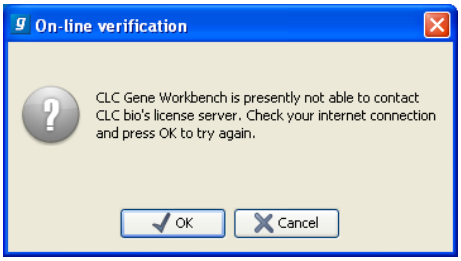

On-line verification
CLC Gene Workbench is presently not able to contact CLC bio's license server. Check your internet connection and press OK to try again.
OK
Cancel

Figure 1.2: *This dialog appears when an online license check is conducted by* **CLC Gene Workbench***, and the computer is off line. Either at start-up or after 24 hours.*

You can then connect to the Internet and retry or you can save your work and close the program. You can run the workbench again later, as long as you are connected to the Internet at start-up.

We use the concept of *"quid quo pro"*. The last two weeks of free demo time given to you is therefore accompanied by a short-form questionnaire where you have the opportunity to give us feedback about the program.

The four weeks demo is offered for each major release of *CLC Gene Workbench*. You will therefore have the opportunity to try the next version (*CLC Gene Workbench 2.0.1*) is released. (If you purchase *CLC Gene Workbench* the first year of updates is included)

1.4.2 Getting and activating the demo license

When you start the program for the first time, you will be presented with the dialog shown in figure 1.3.

If you connect to the internet via a proxy server, click the proxy settings button. Otherwise, just click the "Request evaluation license" button in order to get a license key for a demo of *CLC*

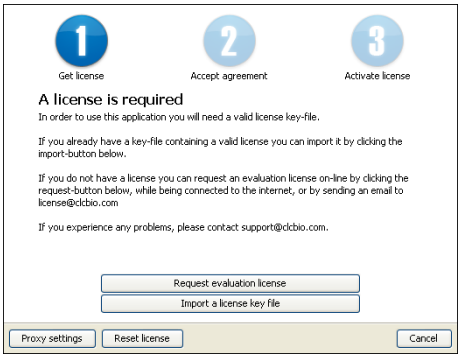

Get license
Accept agreement
Activate license
A license is required
In order to use this application you will need a valid license key-file.
If you already have a key-file containing a valid license you can import it by clicking the import-button below.
If you do not have a license you can request an evaluation license on-line by clicking the request-button below, while being connected to the internet, or by sending an email to license@clcbio.com
If you experience any problems, please contact support@clcbio.com.
Request evaluation license
Import a license key file
Proxy settings
Reset license
Cancel

Figure 1.3: *Selecting "Request evaluation license".*

Gene Workbench 2.0.

Now, our server will issue an evaluation license. This process might take a while depending on your internet connection. When the license key is received, you will be asked to accept the **License agreement** shown in figure 1.4.

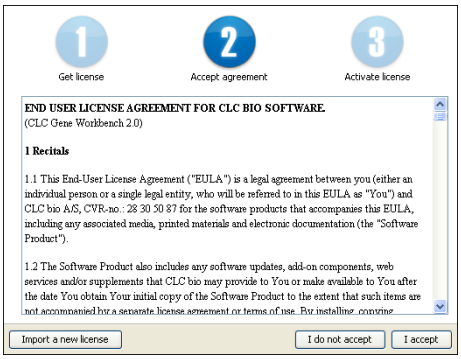

Get license
Accept agreement
Activate license
END USER LICENSE AGREEMENT FOR CLC BIO SOFTWARE
(CLC Gene Workbench 2.0)
1 Recitals
1.1 This End-User License Agreement ("EULA") is a legal agreement between you (either an individual person or a single legal entity, who will be referred to in this EULA as "You") and CLC bio A/S, CVR-no.: 28 30 50 87 for the software products that accompanies this EULA, including any associated media, printed materials and electronic documentation (the "Software Product").
1.2 The Software Product also includes any software updates, add-on components, web services and/or supplements that CLC bio may provide to You or make available to You after the date You obtain Your initial copy of the Software Product to the extent that such items are
Import a new license
I do not accept
I accept

Figure 1.4: *License Agreement.*

Please read the License agreement carefully before clicking **I accept**. In the next step shown in figure 1.5, select "Activate license on-line". Again, you might have to wait for a short while, because the license key is being activated on our server. A license is related to a specific computer, and therefore it can be used by anyone using that computer.

Like in figure 1.3 you can specify a proxy server if needed.

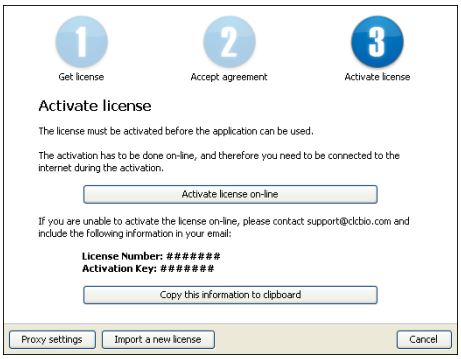

Get license
Accept agreement
Activate license
Activate license
The license must be activated before the application can be used.
The activation has to be done on-line, and therefore you need to be connected to the internet during the activation.
Activate license on-line
If you are unable to activate the license on-line, please contact support@clcbio.com and include the following information in your email:
License Number: #######
Activation Key: #######
Copy this information to clipboard
Proxy settings
Import a new license
Cancel

Figure 1.5: *Activate the license key online.*

Now the license key is activated on your computer, and *CLC Gene Workbench 2.0* starts.

Problems with online activation

If you have problems activating the license online, *CLC Gene Workbench* also offers you an opportunity to manually activating your license key.
Step 3 of the license activation dialog provide a **License number** and an **Activation Key**. By clicking **Copy this information to the clipboard** you can open an email editor and paste these two numbers into the mail. If you email this content and a short explanation to support@clcbio.com we will send back a pre-activated license key.

Also, in all steps of the license dialog you have an option of resetting the license. This will allow you to start over, importing another license. However, information about which licenses were used on the computer is stored externally to prevent unauthorized use of demo licenses.

1.4.3 Commercial license

Unlike the demo version, the commercial version is fully functional offline. When you buy a license for *CLC Gene Workbench*, we will provide you with a license key which is activated as described here.

Start the program, and the dialog shown in figure 1.6 will appear:

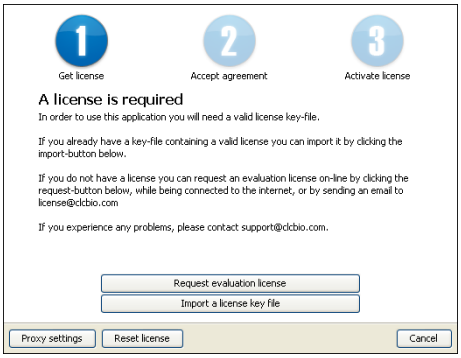

1
Get license
2
Accept agreement
3
Activate license
A license is required
In order to use this application you will need a valid license key-file.
If you already have a key-file containing a valid license you can import it by clicking the import-button below.
If you do not have a license you can request an evaluation license on-line by clicking the request-button below, while being connected to the internet, or by sending an email to license@clcbio.com
If you experience any problems, please contact support@clcbio.com.
Request evaluation license
Import a license key file
Proxy settings
Reset license
Cancel

Figure 1.6: *Select "Import a license key file".*

Choose the option "Import a license key file" in order to specify where your license key is located. Select the license key file provided by CLC bio. When you have selected this file, the **License Agreement** is shown (see figure 1.7). If you want to use another license key instead, click the "Import a license key file" button.

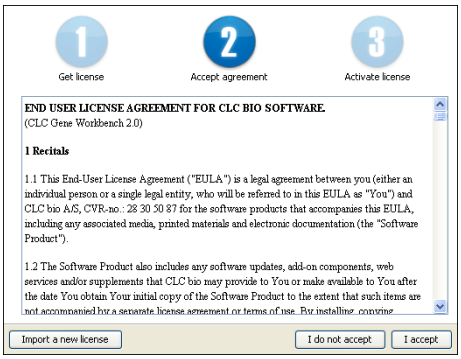

1
Get license
2
Accept agreement
3
Activate license
END USER LICENSE AGREEMENT FOR CLC BIO SOFTWARE.
(CLC Gene Workbench 2.0)
1 Recitals
1.1 This End-User License Agreement ("EULA") is a legal agreement between you (either an individual person or a single legal entity, who will be referred to in this EULA as "You") and CLC bio A/S, CVR-no.: 28 30 50 87 for the software products that accompanies this EULA, including any associated media, printed materials and electronic documentation (the "Software Product").
1.2 The Software Product also includes any software updates, add-on components, web services and/or supplements that CLC bio may provide to You or make available to You after the date You obtain Your initial copy of the Software Product to the extent that such items are
Import a new license
I do not accept
I accept

Figure 1.7: *Read the License Agreement carefully.*

Read the **License Agreement** carefully before clicking the "I accept" button. In the next step shown in figure 1.8, click the "Activate license on-line" button. Your computer must be connected to the internet in order to activate the license. Once the license is activated, you may work

off-line. It will take a little time to activate the license key. When the license key is activated, *CLC Gene Workbench 2.0* will start.

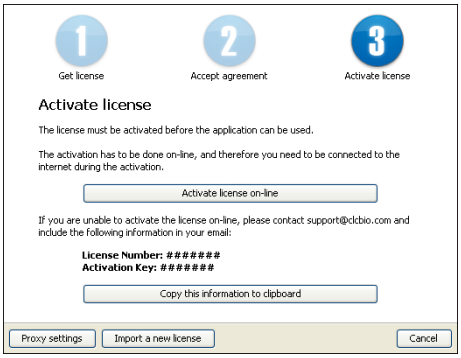

1
Get license
2
Accept agreement
3
Activate license
Activate license
The license must be activated before the application can be used.
The activation has to be done on-line, and therefore you need to be connected to the internet during the activation.
Activate license on-line
If you are unable to activate the license on-line, please contact support@clcbio.com and include the following information in your email:
License Number: #######
Activation Key: #######
Copy this information to clipboard
Proxy settings
Import a new license
Cancel

Figure 1.8: *Activate the license key online.*

A license is related to a specific computer, and therefore it can be used by anyone using that computer. If at some time you want to transfer the license to another computer, please contact license@clcbio.com.

Problems with online activation
If you have problems activating the license online, *CLC Gene Workbench* also offers you an opportunity to manually activating your license key.
Step 3 of the license activation dialog provide a **License number** and an **Activation Key**. By clicking **Copy this information to the clipboard** you can open an email editor and paste these two numbers into the mail. If you email this content and a short explanation to support@clcbio.com we will send back a pre-activated license key.

Also, in all steps of the license dialog you have an option of resetting the license. This will allow you to start over, importing another license. However, information about which licenses were used on the computer is stored externally to prevent unauthorized use of demo licenses.

1.4.4 Upgrading from a demo license to a commercial license

If you are trying a demo of *CLC Gene Workbench*and want to upgrade to a license that you have bought, choose **Upgrade license** in the Help menu. Then follow the description in section 1.4.3.

1.5 About CLC Workbenches

In November 2005 CLC bio released two Workbenches: *CLC Free Workbench* and *CLC Protein Workbench*. *CLC Protein Workbench* is developed from the free version, giving it the well-tested user friendliness and look & feel. However, the *CLC Protein Workbench* includes a range of more advanced analyses.
In March 2006, *CLC Gene Workbench* and *CLC Combined Workbench* were added to the product portfolio of CLC bio. Like *CLC Protein Workbench*, *CLC Gene Workbench* builds on *CLC Free Workbench*. It shares some of the advanced product features of *CLC Protein Workbench*, and it has additional advanced features. *CLC Combined Workbench* holds all basic and advanced features of the *CLC Workbenches*.
For an overview of which features the four workbenches include, see `http://www.clcbio.com/features`.

All workbenches will be improved continuously. If you have a CLC Free Workbench or a commercial workbench, and you are interested in receiving news about updates, you should register your e-mail and contact data on http://www.clcbio.com, if you haven't already registered when you downloaded the program.

1.5.1 New program feature request

The CLC team is continuously improving the program with our users' interest in mind. Therefore, we welcome all requests from users, and they can be submitted from our homepage http://www.clcbio.com. Likewise, you are more than welcome to suggest new features or more general improvements to the program on support@clcbio.com.

1.5.2 Report program errors

CLC bio is doing everything possible to eliminate program errors. Nevertheless, some errors might have escaped our attention. If you discover an error in the program, you can use the **Report a Program Error** function in the **Help** menu of the program to report it. In the **Report a Program Error** dialog you are asked to write your e-mail address. This is because we would like to be able to contact you for further information about the error or for helping you with the problem.

Notice that no personal information is send via the error report. Only the information which can be seen in the **Program Error Submission Dialog** is submitted.

You can also write an e-mail to support@clcbio.com. Remember to specify how the program error can be reproduced.

All errors will be treated seriously and with gratitude.

We appreciate your help.

Start in safe mode

If the program becomes unstable on start-up, you can start it in **Safe mode**. This is done be pressing down the Shift button while the program starts.

When starting in safe mode, the user settings (e.g. the settings in the **Side Panel**) are deleted and cannot be restored. Your data stored in the **Navigation Area** is not deleted.

1.5.3 Free vs. commercial workbenches

The advanced analyses of the commercial workbenches, *CLC Protein Workbench* and *CLC Gene Workbench* are not present in *CLC Free Workbench*. Likewise, some advanced analyses are available in *CLC Gene Workbench* but not in *CLC Protein Workbench*, and visa versa. All types of basic and advanced analyses are available in *CLC Combined Workbench*.
However, the output of the commercial workbenches can be viewed in all other workbenches. This allows you to share the result of your advanced analyses from e.g. *CLC Combined Workbench*, with people working with e.g. *CLC Free Workbench*. They will be able to view the results of your analyses, but not redo the analyses.

The CLC Workbenches are developed for Windows, Mac and Linux platforms. Data can be

exported/imported between the different platforms in the same easy way as when exporting/importing between two computers with e.g. Windows.

This is illustrated in figure 1.9.

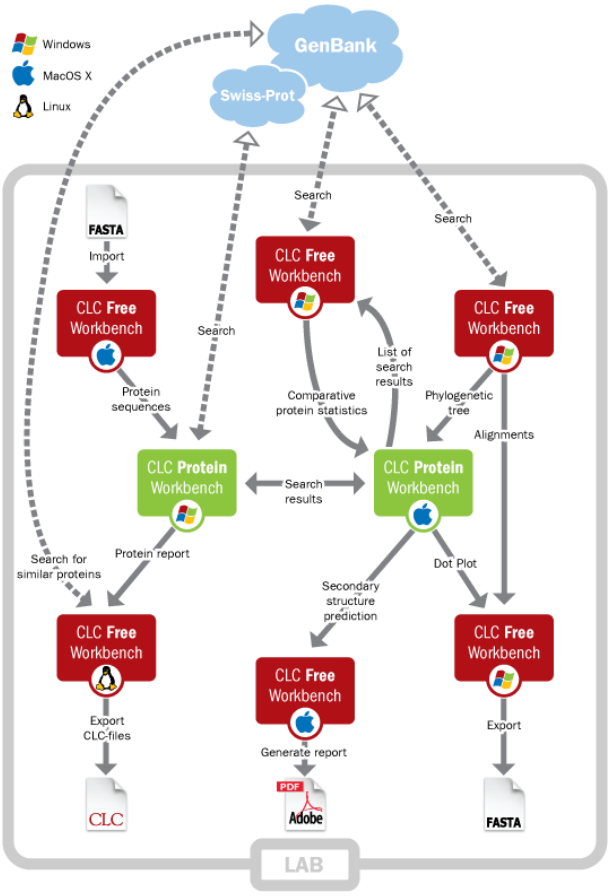


Figure 1.9: *An example of how research can be organized and how data can flow between users of different workbenches, working on different platforms.*

1.6 When the program is installed: Getting started

CLC Gene Workbench 2.0 includes an extensive **Help** function, which can be found in the **Help** menu of the program's **Menu bar**. The **Help** function can also be launched by pressing F1. The help topics are sorted in a table of contents and the topics can be searched.

1.6.1 Basic concepts of using CLC Workbenches

Here is a short list of basic concepts of how to use *CLC Gene Workbench*:

- All data for use in the *CLC Gene Workbench* should be stored inside the program in the **Navigation Area**. This means that you have to either import some of your own data or use e.g. the GenBank search function ( ).
- The data can be viewed in a number of ways. First, click the element (e.g. a sequence) in the **Navigation Area** and then click **Show**( ) to find a proper way to view the data (see figure 1.10 for an example).

- When a view is opened, there are three basic ways of interacting:
  1. Using the **Side Panel** to the right to specify how the data should be displayed (these settings are not associated with your data but they can be saved by clicking the icon ( ) in the upper right corner of the **Side Panel**).
  2. Using right-click menus e.g. to edit a sequence (in this case you have to make a selection first using the selection mode( )).
  3. Using the Zoom ( ) / ( ) tools.
- In the Toolbox, you find all the tools for analyzing and working on your data. In order to use these tools, your data must be stored in a project in the **Navigation Area**

Figure 1.10: *The different ways of viewing DNA sequences.*

1.6.2 Quick start

When the program opens for the first time, the background of the workspace is visible. In the background are three quick start shortcuts, which will help you getting started. These can be seen in figure 1.11.

Figure 1.11: *Three available Quick start short cuts, available in the background of the workspace.*

The function of the three quick start shortcuts is explained here:

- **Import data.** Opens the **Import** dialog, which you let you browse for, and import data from your file system.
- **New sequence.** Opens a dialog which allows you to enter your own sequence.
- **Read tutorials.** Opens the tutorials a menu with a number of tutorials. These are also available from the **Help** menu in the **Menu bar**.

It might be easier to understand the logic of the program by trying to do simple operations on existing data. Therefore *CLC Gene Workbench 2.0* includes an example data set, which can be found on our web page, or downloaded from the program (Also found in the **Help** menu).

1.6.3 Import of example data

When downloading *CLC Gene Workbench 2.0* you are asked if you would like to import an example data set. If you accept, the data is downloaded automatically and saved in the program. If you didn't download the data, or for some other reason need to download the data again, you have two options.

You can click ( ) **Install example data** in the **Help** menu of the program. This installs the data automatically. You can also go to our website at `http://www.clcbio.com`, Software/CLC Free Workbench/Example data , and download the example data from there.

If you download the file from the website, you need to import it into the program. See chapter 6.1 for more about importing data.

1.7 Network configuration

If you use a proxy server to access the Internet you must configure *CLC Gene Workbench 2.0* to use this. Otherwise you will not be able to perform any on-line activities (e.g. searching GenBank). *CLC Gene Workbench 2.0* supports the use of a HTTP-proxy and an anonymous SOCKS-proxy.

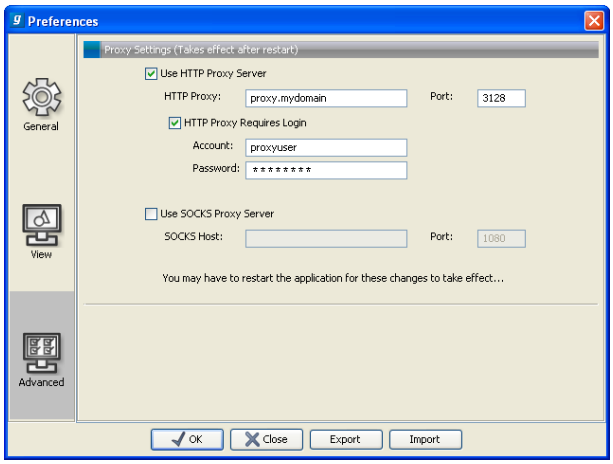


Figure 1.12: *Adjusting proxy preferences.*

To configure your proxy settings, open *CLC Gene Workbench 2.0*, and go to the **Advanced**-tab of the **Preferences** dialog (figure 1.12) and enter the appropriate information.
You have the choice between a HTTP-proxy and a SOCKS-proxy. *CLC Gene Workbench 2.0* only supports the use of a SOCKS-proxy that does not require authorization.

If you have any problems with these settings you should contact your systems administrator.

1.8 Adjusting the maximum amount of memory

If you have a large amount of memory (RAM) available in your system and need to work with very large data objects, you can manually change the maximum amount of memory available to the program. Doing so is a somewhat complicated, unsupported procedure and may cause the program to fail if done incorrectly.

Depending on your operating system you may have to repeat these changes if you update *CLC Gene Workbench 2.0* to a newer version.

1.8.1 Microsoft Windows

- Locate the *CLC Gene Workbench 2.0* directory inside your Program Files directory and open it
- Create a new, empty text-file called clcwb.vmoptions (make sure the filename does not end with ".txt")
- Add a single line to the file with a syntax similar to:

  ```
  -Xmx512m
  ```

  It is very important that the line looks exactly like the one in the example above, and that you only change the value of the number (512 in the example). For the best performance you should not choose a number greater than the amount (in megabytes) of physical memory available on your system.

1.8.2 Mac OS X

- Locate the CLC Free Workbench program file in your Applications folder
- Right-click / control-click the file and choose "Show Package Contents" from the pop-up menu
- Open the file called "Info.plist" located inside the "Contents" folder using the "Property List Editor" application or a text editor like "TextEdit"
- Edit the Root/Java/VMOptions property, and set the maximum amount of memory to a desired value. The property has a specific syntax similar to:

  ```
  -Xmx512m
  ```

  It is very important that you only change the value of the number, 512 in the example above, to the amount of megabytes you want. For the best performance you should not choose a number greater than the amount of physical memory available on your system.

1.8.3 Linux

- Locate the directory where you installed *CLC Gene Workbench 2.0* and open it.
- Create a new, empty text-file called "clcwb.vmoptions".
- Add a single line to the file with a syntax similar to:

  ```
  -Xmx512m
  ```

  It is very important that the line looks exactly like the one in the example above, and that you only change the value of the number (512 in the example). For the best performance you should not choose a number greater than the amount (in megabytes) of physical memory available on your system.

1.9 The format of the user manual

This user manual offers support to Windows, Mac OS X and Linux users. The software is very similar on these operating systems. In areas where differences exist, these will be described separately. However, the term "right-click" is used throughout the manual, but some Mac users may have to use Ctrl+click in order to perform a "right-click" (if they have a single-button mouse).

The most recent version of the user manuals can be downloaded from `http://www.clcbio.com/usermanuals`.

The user manual consists of four parts.

- The **first part** includes the introduction and some tutorials showing how to apply the most significant functionalities of *CLC Gene Workbench 2.0*.
- The **second part** describes in detail how to operate all the program's basic functionalities.
- The **third part** digs deeper into some of the bioinformatic features of the program. In this part, you will also find our "Bioinformatics explained" sections. These sections elaborate on the algorithms and analyses of *CLC Gene Workbench 2.0* and provide more general knowledge of bioinformatic concepts.
- The **fourth part** is the Appendix and Index.

Each chapter includes a short table of contents.

1.9.1 Text formats

In order to produce a clearly laid-out content in this manual, different formats are applied:

- A feature in the program is in bold starting with capital letters. (Example: **Navigation Area**)
- An explanation of how a particular function is activated, is illustrated by "|" and bold. (E.g.: **select the element** | **Edit** | **Rename**)
- Icons, such as "( )", are included in order to ease the navigation in the **Toolbox**.
- The format of the program name is bold and italic: *CLC Gene Workbench 2.0*
- The captions of displayed screenshots are in *italic*.

Chapter 2

Tutorials

Contents

2.1 Tutorial: Starting up the program . . . 26
2.1.1 Creating a project and a folder . . . 26
2.1.2 Import data . . . 27
2.1.3 Supported data formats . . . 28
2.2 Tutorial: View sequence . . . 29
2.3 Tutorial: GenBank search and download . . . 31
2.3.1 Saving the search . . . 32
2.3.2 Searching for matching objects . . . 32
2.3.3 Saving the sequence . . . 32
2.4 Tutorial: Align protein sequences . . . 32
2.4.1 Alignment dialog . . . 33
2.5 Tutorial: Create and modify a phylogenetic tree . . . 34
2.5.1 Tree layout . . . 35
2.6 Tutorial: Detect restriction sites . . . 35
2.6.1 View restriction site . . . 36
2.7 Tutorial: Sequence information . . . 37
2.8 Tutorial: BLAST search . . . 38
2.9 Tutorial: Primer design . . . 41
2.9.1 Finding the region to amplify . . . 41
2.9.2 Specifying a region for the forward primer . . . 41
2.9.3 Examining the primer suggestions . . . 42
2.9.4 Calculating a primer pair . . . 44
2.10 Tutorial: Assembly . . . 45
2.10.1 Getting an overview of the contig . . . 45
2.10.2 Finding and editing inconsistencies . . . 45
2.10.3 Documenting your changes . . . 46
2.11 Tips and tricks for the experienced user . . . 47
2.11.1 Open and arrange views using drag and drop . . . 48
2.11.2 Find element in the Navigation Area . . . 48
2.11.3 Find specific annotations on a sequence . . . 49

2.11.4 Split sequences into several lines 50
2.11.5 Make a new sequence of a coding region 50
2.11.6 Translate a coding region 50
2.11.7 Copy annotations from one sequence to another 51
2.11.8 Get overview and detail of a sequence at the same time 51
2.11.9 Smart selecting in sequences and alignments 52
2.11.10 Check for updates and additional information about sequences 52
2.11.11 Quickly import sequences using copy-paste 53
2.11.12 Perform analyses on many elements 53
2.11.13 Drag elements to the Toolbox 54
2.11.14 Export elements while preserving history 54
2.11.15 Avoid the mouse trap - use keyboard shortcuts 55

This chapter contains tutorials representing some of the features of *CLC Gene Workbench 2.0*. The first tutorials are meant as a short introduction to operating the program. The last tutorials give examples of how to use some of the main features of *CLC Gene Workbench 2.0*.
The tutorials are also available as interactive Flash tutorials on `http://www.clcbio.com/tutorials`.

2.1 Tutorial: Starting up the program

This brief tutorial will take you through the most basic steps of working with *CLC Gene Workbench*. The tutorial introduces the user interface, demonstrates how to create a project, and demonstrates how to import your own existing data into the program.

When you open *CLC Gene Workbench* for the first time, the user interface looks like figure 2.1.

At this stage, the important issues are the **Navigation Area** and the **View Area**.

The **Navigation Area** to the left is where you keep all your data for use in the program. Most analyses of *CLC Gene Workbench* require that the data is saved in the **Navigation Area**. There are several ways to get data into the **Navigation Area**, and this tutorial describes how to import existing data.

The **View Area** is the main area to the right. This is where the data can be 'viewed'. In general, a **View** is a display of a piece of data, and the **View Area** can include several **Views**. The **Views** are represented by tabs, and can be organized e.g. by using 'drag and drop'.

2.1.1 Creating a project and a folder

When *CLC Gene Workbench* is started there is one default project in the **Navigation Area**. Create an additional project by:

File in the Menu Bar | **New** | **Project (** **)**
or **Ctrl + R (⌘ + R on Mac)**

Name the project 'Test' and press **Enter**.

The data in the project can be further organized into folders. Create a folder in the 'Test' project by:

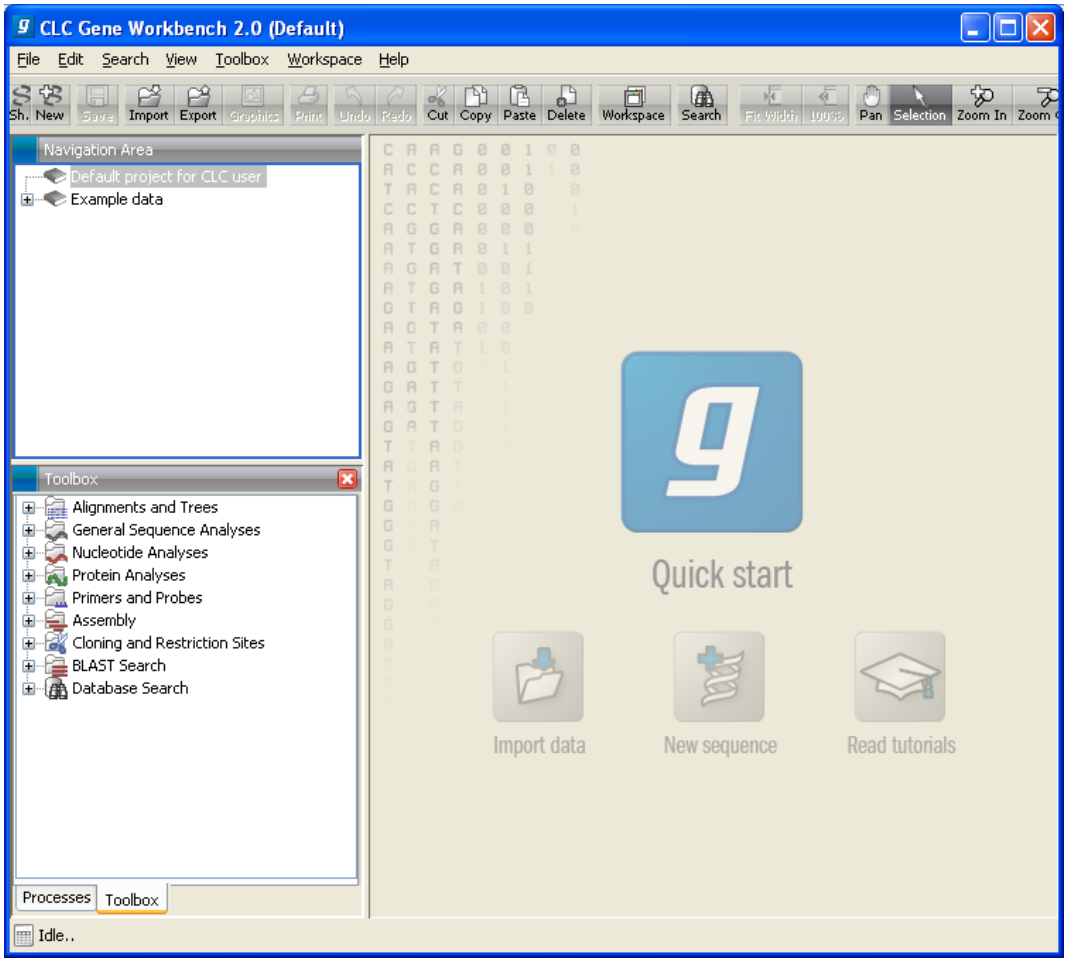


Figure 2.1: *The user interface as it looks when you start the program for the first time. (Windows version of* **CLC Gene Workbench***. The interface is similar for Mac and Linux.)*

Right-click the 'Test'-project in the Navigation Area | New | Folder ( )
or **Ctrl + F (⌘ + F on Mac)**

Name the folder 'Subfolder' and press **Enter**.

2.1.2 Import data

Next, we want to import a sequence called HUMDINUC.fsa (FASTA format) from our own Desktop into the new 'Subfolder'. (This file is chosen for demonstration purposes only - you may have another file on your desktop, which you can use to follow this tutorial. You can import all kinds of files.)

In order to import the HUMDINUC.fsa file:

Import ( ) in the Toolbar | select FASTA (.fsa/.fasta) in the (Files of type) drop down menu | navigate to HUMDINUC.fsa on the desktop | Select

For files of FASTA or PIR format, you are asked to state which type of sequence you are importing. (This will ensure that *CLC Gene Workbench* treats the sequence in the correct way.)

Click DNA/RNA | OK

The sequence is imported into the project or folder that was selected in the **Naviagation Area**, before you clicked **Import**. Double-click the sequence in the **Navigation Area** to view it. The final result looks like figure 2.2.

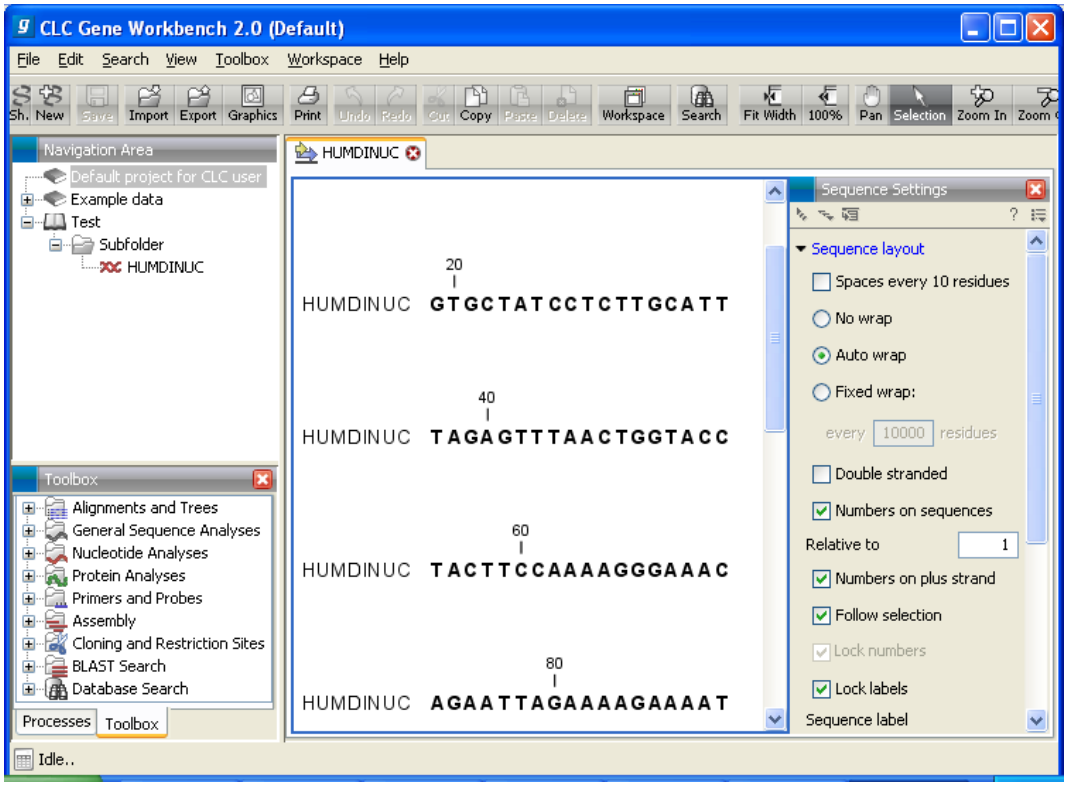


Figure 2.2: *The HUMDINUC file is imported and opened.*

2.1.3 Supported data formats

CLC Gene Workbench can import and export the following formats:

File type	Suffix	File format used for
Phylip Alignment	.phy	alignments
GCG Alignment	.msf	alignments
Clustal Alignment	.aln	alignments
Newick	.nwk	trees
FASTA	.fsa/.fasta	sequences
GenBank	.gbk/.gb/.gp	sequences
GCG sequence	.gcg	sequences (only import)
PIR (NBRF)	.pir	sequences (only import)
Staden	.sdn	sequences (only import)
VectorNTI		sequences (only import)
DNAstrider	.str/.strider	sequences
Swiss-Prot	.swp	protein sequences
Lasergene sequence	.pro	protein sequence (only import)
Lasergene sequence	.seq	nucleotide sequence (only import)
Embl	.embl	nucleotide sequences
Nexus	.nxs/.nexus	sequences, trees, alignments, and sequence lists
CLC	.clc	sequences, trees, alignments, reports, etc.
Text	.txt	all data in a textual format
ABI		Trace files (only import)
AB1		Trace files (only import)
SCF2		Trace files (only import)
SCF3		Trace files (only import)
Phred		Trace files (only import)
mmCIF	.cif	structure (only import)
PDB	.pdb	structure (only import)
Preferences	.cpf	CLC workbench preferences

Notice that *CLC Gene Workbench* can import 'external' files, too. This means that *CLC Gene Workbench* can import all files and display them in the **Navigation Area**, while the above mentioned formats are the types which can be read by *CLC Gene Workbench*.

2.2 Tutorial: View sequence

This brief tutorial will take you through some different ways to display a sequence in the program. The tutorial introduces zooming on a sequence, dragging tabs, and opening selection in new view.

We will be working with DNA sequence 'AY738615'. Double-click the sequence in the **Navigation Area** to open it. The sequence is displayed with annotations above it. (To provide a better view of the sequence, hide the **Side Panel**. This is done by clicking the red X ( ) at the top right corner of the **Side Panel** (in the right side of the **View Area**). (See figure 2.3).

As default, *CLC Gene Workbench* displays a sequence with annotations (colored arrows on the sequence) and zoomed to see the residues.

In this tutorial we want to have an overview of the whole sequence. Hence;

click Zoom Out ( ) in the Toolbar | click the sequence until you can see the whole sequence

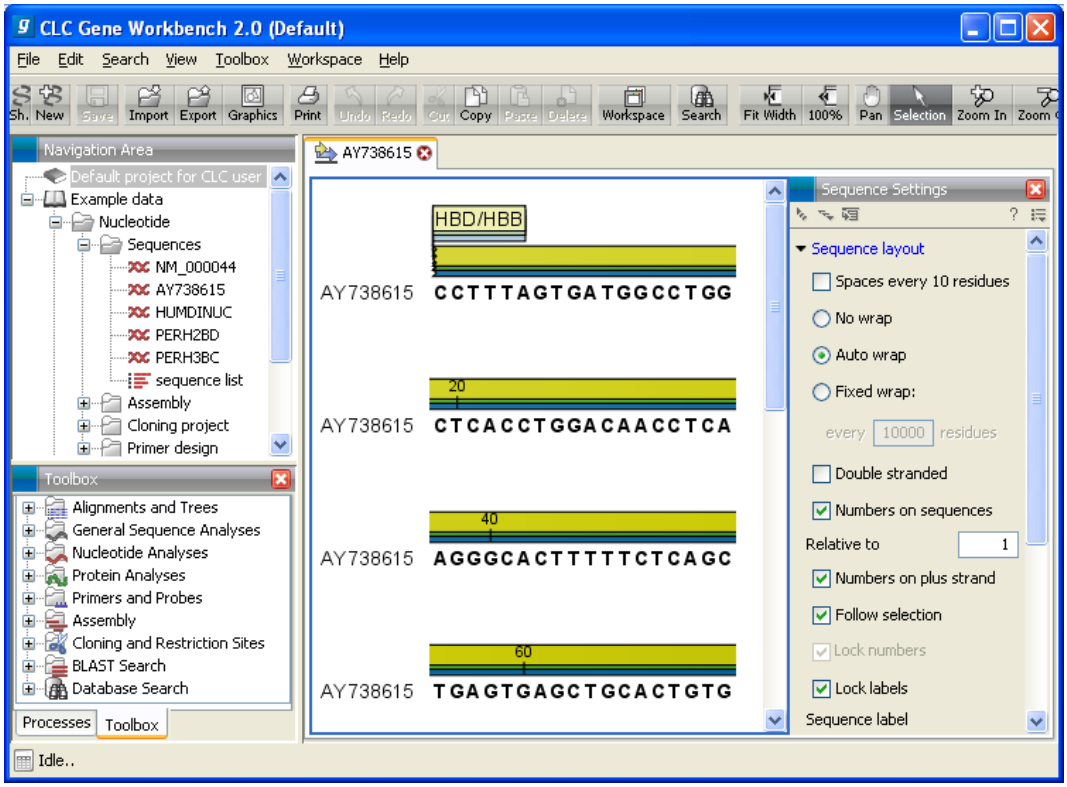


Figure 2.3: *DNA sequence 'AY738615' opened in a view. The view preferences has been hidden to provide more space for the view.*

In the following we will show how the same sequence can be displayed in two different views:

double-click sequence 'AY738615' in the Navigation Area

This opens an additional tab. Drag this tab to the bottom of the view. (See figure 2.4).

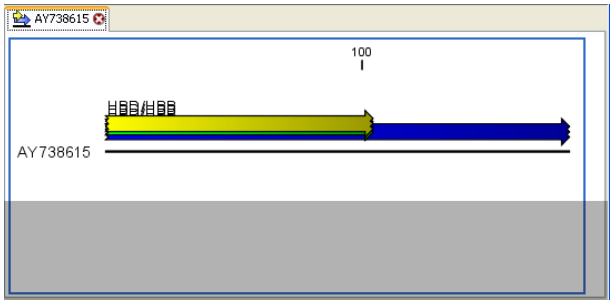


Figure 2.4: *Dragging the tab down to the bottom of the view will display a gray area indicating that the tab can be "dropped" here and split the view.*

The result is two views of the same sequence in the **View Area**, as can be seen in figure 2.5.

If you want to display a part of the sequence, it is possible to select it, and open it in another view:

click Selection ( ) in Toolbar | select a part of the sequence | right-click the selected part of the sequence in the top view | Open Selection in New View

This opens a third display of sequence 'AY738615'. However, only the part which was selected. In order to make room for displaying the selection of the sequence (the most recent view), drag

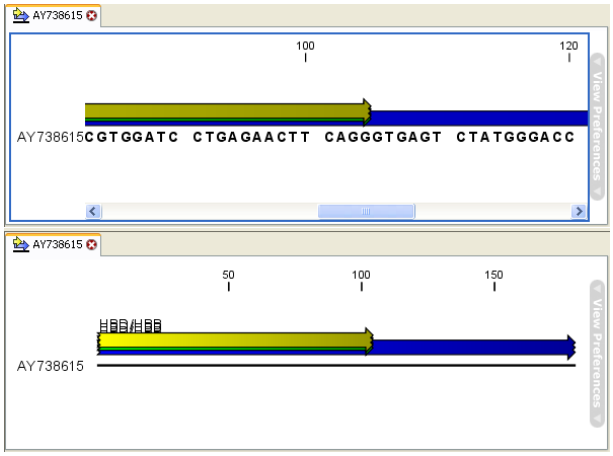


Figure 2.5: *The resulting two views which are split horizontally.*

the tab of the view down, next to the tab of the bottom view.

2.3 Tutorial: GenBank search and download

The *CLC Gene Workbench* allows you to search the NCBI GenBank database directly from the program, giving you the opportunity to both open, view, analyze and save the search results without using any other applications. To conduct a search in NCBI GenBank from *CLC Gene Workbench* you must be connected to the Internet.

This tutorial shows how to find a complete human hemoglobin DNA sequence in a situation where you do not know the accession number of the sequence.

To start the search:

Search | Search NCBI Entrez ( )

This opens the search view. We are searching for a DNA sequence, hence:

Nucleotide

Now we are going to **Adjust Parameters** for the search. By clicking **More Choices** you activate an additional set of fields where you can enter search criteria. Each search criterion consists of a drop down menu and a text field. In the drop down menu you choose which part of the NCBI database to search, and in the text field you enter what to search for:

Click More Choices until three search criteria are available | choose Organism in the first drop down menu | write 'human' in the adjoining text field | choose All Fields in the second drop down menu | write 'hemoglobin' in the adjoining text field | choose All Fields in the third drop down menu | write 'complete' in the adjoining text field

Now you have two choices: Either to click **Start search( )** to commence the search in NCBI, or to click **Save search parameters( )** to choose where to save the search.

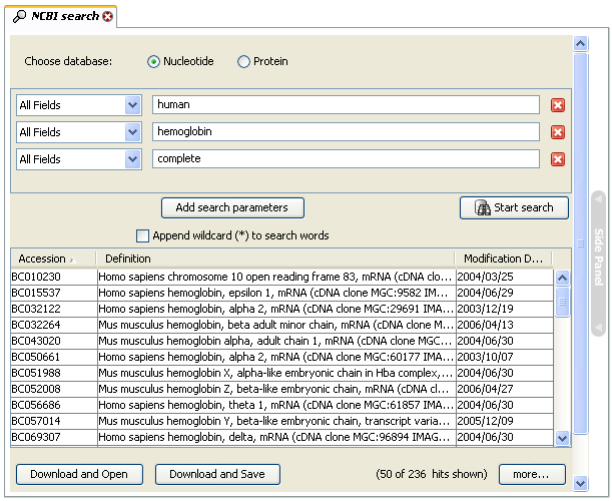

NCBI search
Choose database:
Nucleotide
Protein
All Fields
human
All Fields
hemoglobin
All Fields
complete
Add search parameters
Start search
Append wildcard (*) to search words
Accession
Definition
Modification D...
BC010230 Homo sapiens chromosome 10 open reading frame 83, mRNA (cDNA clo... 2004/03/25
BC015537 Homo sapiens hemoglobin, epsilon 1, mRNA (cDNA clone MGC:9582 IM... 2004/06/29
BC032122 Homo sapiens hemoglobin, alpha 2, mRNA (cDNA clone MGC:29691 IMA... 2003/12/19
BC032264 Mus musculus hemoglobin, beta adult minor chain, mRNA (cDNA clone M... 2006/04/13
BC043020 Mus musculus hemoglobin alpha, adult chain 1, mRNA (cDNA clone MGC... 2004/06/30
BC050661 Homo sapiens hemoglobin, alpha 2, mRNA (cDNA clone MGC:60177 IMA... 2003/10/07
BC051988 Mus musculus hemoglobin X, alpha-like embryonic chain in Hba complex,... 2004/06/30
BC052008 Mus musculus hemoglobin Z, beta-like embryonic chain, mRNA (cDNA cl... 2006/04/27
BC056686 Homo sapiens hemoglobin, theta 1, mRNA (cDNA clone MGC:61857 IMA... 2004/06/30
BC057014 Mus musculus hemoglobin Y, beta-like embryonic chain, transcript varia... 2005/12/09
BC069307 Homo sapiens hemoglobin, delta, mRNA (cDNA clone MGC:96894 IMAG... 2004/06/30
Download and Open
Download and Save
(50 of 236 hits shown)
more...

Figure 2.6: *NCBI search view.*

2.3.1 Saving the search

If you click 'Save search parameters', the program does not save the search results, but rather the search criteria. This allows you to perform exactly the same search later on.

In this tutorial, we are not certain of the quality of our search criteria, and therefore we choose not to save them. Consequently, click **Start search(** **)** to perform the search.

2.3.2 Searching for matching objects

When the search is complete, the list of hits is shown. If the desired complete human hemoglobin DNA sequence is found, the sequence can be viewed by double-clicking it in the list of hits from the search. If the desired sequence is not shown, you can click the 'More' button below the list to see more hits.

2.3.3 Saving the sequence

The sequences which are found during the search can be displayed by double-clicking in the list of hits. However, this does not save the sequence. It is necessary to save the sequences before any analysis can be conducted. A sequence is saved like this:

click the tab with the name of the sequence | Save in the toolbar ( )

or **click the tab with the name of the sequence | Ctrl + S (⌘ + S on Mac)**

When you close the view of the sequence, you are asked if you want to save the file.

If you do not want to view the sequence first, the sequence can be saved by dragging it from the list of hits into the **Navigation Area**.

2.4 Tutorial: Align protein sequences

It is possible to create multiple alignments of nucleotide and protein sequences. *CLC Gene Workbench* offers several opportunities to view alignments. The alignments can be used for building phylogenetic trees.

The sequences must be saved in the **Navigation Area** in order to be included in an alignment. To save a sequence which is displayed in the **View Area**, click the tab of the sequence and press Ctrl + S (or ⌘ + S on Mac). In this tutorial eight protein sequences from the Example data will be aligned. (See figure 2.7).

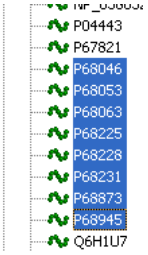


Figure 2.7: *Eight protein sequences in a Protein project in the Navigation Area.*

To begin aligning the protein sequences:

> **select the sequences** | **right-click either of the sequences** | **Toolbox** | **Alignments and Trees(**) | **Create Alignment (**)

2.4.1 Alignment dialog

This opens the dialog shown in fig. 2.8.

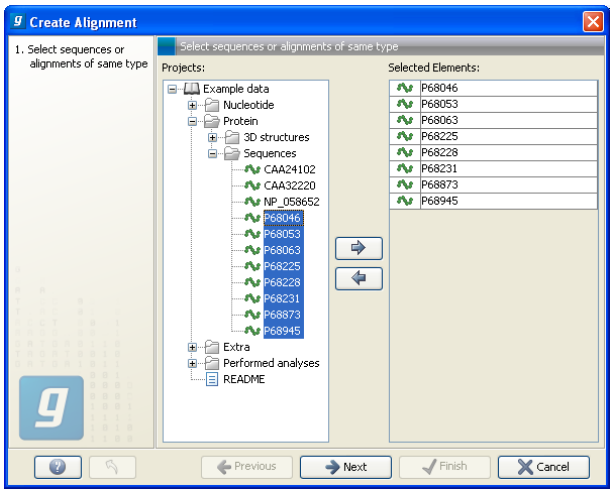


Figure 2.8: *The alignment dialog displaying the 8 chosen protein sequences.*

It is possible to add and remove sequences from **Selected Elements** list. When the relevant proteins are selected there are two options: Click **Next** to adjust parameters for the alignment.

Clicking **Next** opens the dialog shown in fig. 2.9.

Leave the parameters at their default settings. An explanation of the parameters can be found in the program's **Help** function ( ) or in the user manual on `http://www.clcbio.com/download`.

Click **Finish** to start the alignment process which is shown in the **Toolbox** under the **Processes** tab. When the program is finished calculating it displays the alignment (see fig. 2.10):

Notice! The new alignment is not saved automatically. (The text on the tab is bold and italic to illustrate this.)

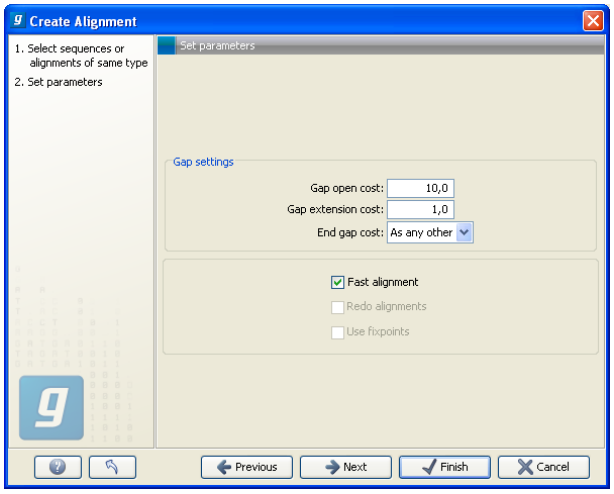


Figure 2.9: *The alignment dialog displaying the available parameters which can be adjusted.*

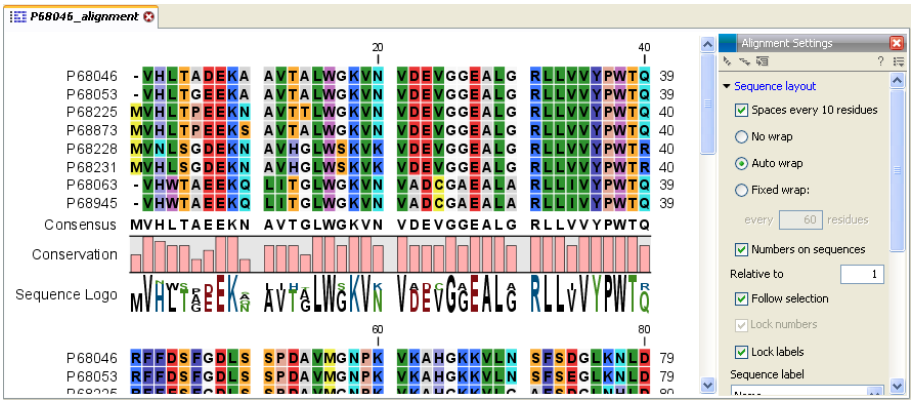


Figure 2.10: *The resulting alignment.*

To save the alignment, drag the tab of the alignment view into the **Navigation Area**.

2.5 Tutorial: Create and modify a phylogenetic tree

You can make a phylogenetic tree from an existing alignment. (See how to create an alignment in 'Tutorial: Align protein sequence').

We use the 'P04443_alignment' located in Performed Analyses: Protein Workbench in the Example data. To create a phylogonetic tree:

> **right-click the 'P04443_alignment' in the Navigation Area | Toolbox | Alignments and Trees( ) | Create Tree (:)**

A dialog opens where you can confirm your selection of the alignment. Moving to the next step in the dialog you can choose between the neighbor joining and the UPGMA algorithms for making trees. You also have the option of including a bootstrap analysis of the result.

Click **Finish** to start the calculation, which can be seen in the **Toolbox** under the **Processes** tab, and after a short while a tree appears in the **View Area** (figure 2.11).

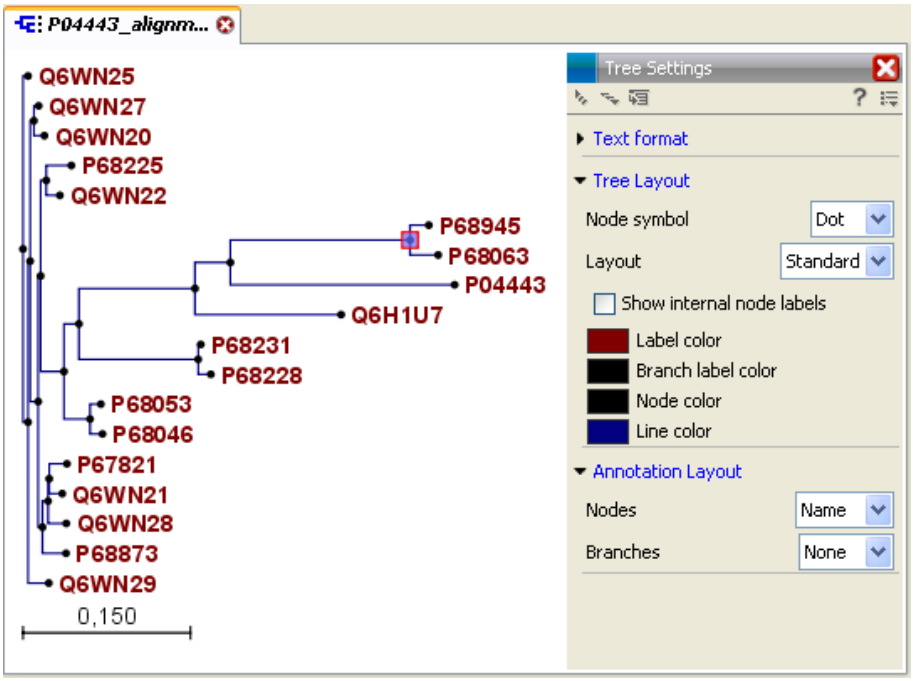


Figure 2.11: *After choosing which algorithm should be used, the tree appears in the View Area. The Side panel in the right side of the view allows you to adjust the way the tree is displayed.*

2.5.1 Tree layout

Using the **View preferences** (in the right side of the interface) of the tree view, you can edit the way the tree is displayed. Click **Tree Layout** and open the **Layout** drop down menu. Here you can choose between standard and topology layout. The topology layout can help to give an overview of the tree if some of the branches are very short.

When the sequences include the appropriate annotation, it is possible to choose between the accession number and the species names at the leaves of the tree. Sequences downloaded from GenBank, for example, have this information. The **Annotation Layout** preferences allows these different node annotations as well as different annotation on the branches.

The branch annotation includes the bootstrap value, if this was selected when the tree was calculated. It is also possible to annotate the branches with their lengths.

2.6 Tutorial: Detect restriction sites

This tutorial will show you how to find restriction sites and annotate them on a sequence.

Suppose you are working with sequence PERH3BC from the example data, (can be downloaded from `http://www.clcbio.com/download`) and you wish to know which restriction enzymes will cut this sequence exactly once and create a 3´ overhang. Do the following:

> **select the PERH3BC sequence from the Primer design folder | Toolbox in the Menu Bar | Cloning and Restriction Sites ( ) | Restriction sites ( )**

The dialog shown in (fig. 2.12) opens, and you can confirm or change your selection of input sequence.

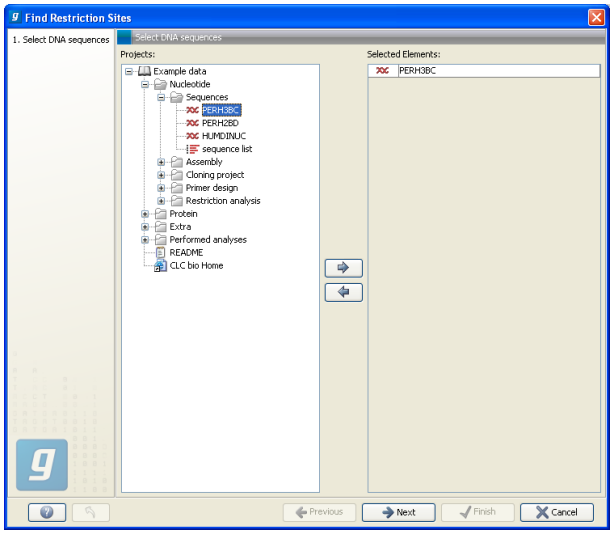


Figure 2.12: *Choosing sequence PERH3BC.*

In the next step you uncheck "Blunt ends" and "5´ overhang" since we only wish to use enzymes with a "3´ overhang". Then click **Select all** (see figure 2.13).

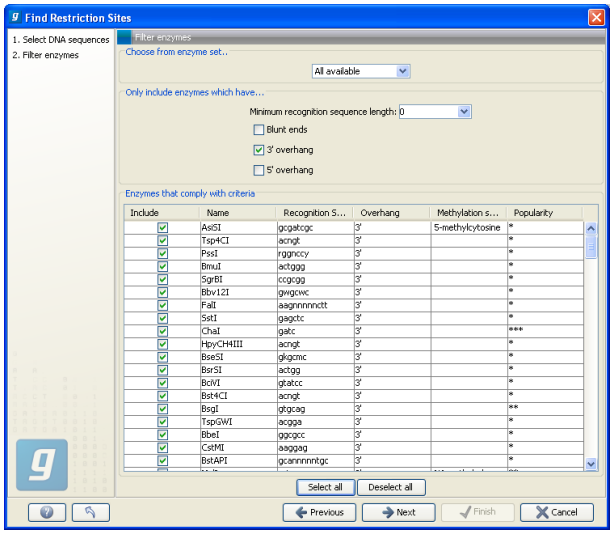


Figure 2.13: *Setting parameters for restriction site detection.*

Click Next and choose both textual and graphical output. (See figure 2.14).

Click **Finish** to start the restriction site analysis.

2.6.1 View restriction site

The restriction sites are shown in two views: one view is in a textual format and the other view displays the sites as annotations on the sequence. To see both views at once:

View in the menu bar | Split Horizontally ( )

The result is shown in figure 2.15.

Notice! The results are not automatically saved.

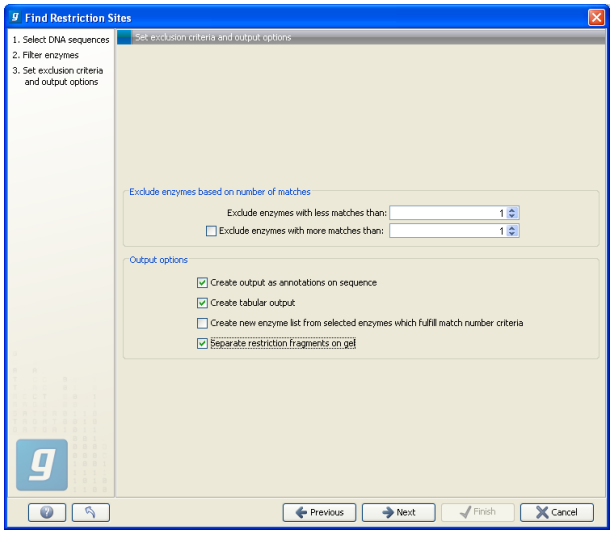

Find Restriction Sites
1. Select DNA sequences
2. Filter enzymes
3. Set exclusion criteria and output options
Set exclusion criteria and output options
Exclude enzymes based on number of matches
Exclude enzymes with less matches than:
Exclude enzymes with more matches than:
Output options
Create output as annotations on sequence
Create tabular output
Create new enzyme list from selected enzymes which fulfill match number criteria
Separate restriction fragments on gel
Previous
Next
Finish
Cancel

Figure 2.14: *Selecting enzymes.*

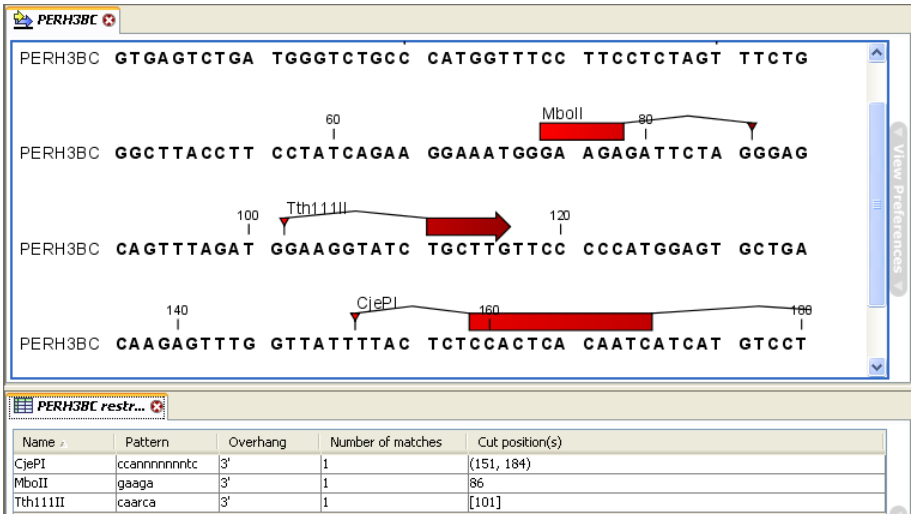


Figure 2.15: *The result of the restriction site detection is displayed as text, and in this tutorial the View shares the View Area with a View of the PERH3BC sequence displaying the restriction sites (split-screen-view).*

To save the result:

Right-click the tab | **File** | **Save (** **)**

2.7 Tutorial: Sequence information

This tutorial shows you how to see background information about a sequence, including an overview of its annotations.

Suppose you are working with the HUMHBB sequence from the example data, (The Example data can be installed in the program by: clicking **Install Example Data** from the **Help** menu in the **Menu Bar**. The Example data can also be downloaded from `http://www.clcbio.com/download`.) and you wish to see more background information about this sequence. This can be done using the **Sequence Info** functionality of *CLC Gene Workbench*:

Select HUMHBB in the Navigation Area | Show ( ) in Menu Bar | Sequence Info ( )

This opens a new view shown in figure 2.16.

Figure 2.16: *The initial view of sequence info of HUMHBB.*

The sequence is originally downloaded from GenBank, and it is the information from the GenBank file which is shown as a list of headings. Click the heading **Modification Date** to see when the sequence was modified in GenBank.

At the bottom there is an **Annotation Map** providing an overview of the annotations on the sequence. The annotations are divided into types. We are interested in the coding sequences of HUMHBB:

Click Annotation Map | Click CDS

The seven coding sequences are displayed with the corresponding positions in GenBank syntax. In order to make full use of the **Annotation Map**, open a normal view of the HUMHBB sequence below the **Sequence Info**:

Select the HUMHBB in the Navigation Area | Drag it to the bottom of the View Area until a gray shadow appears

Now, clicking a coding sequences in the **Annotation Map** will make a selection representing the coding sequence in the view below . You can see that the selection matches the CDS annotation the yellow boxes in figure 2.17).

2.8 Tutorial: BLAST search

This tutorial shows you how to perform a BLAST search using *CLC Gene Workbench*.

Suppose you are working with the NP_058652 protein which constitutes the beta part of the hemoglobin molecule that is expressed in the adult house mouse, *Mus musculus*. To obtain more

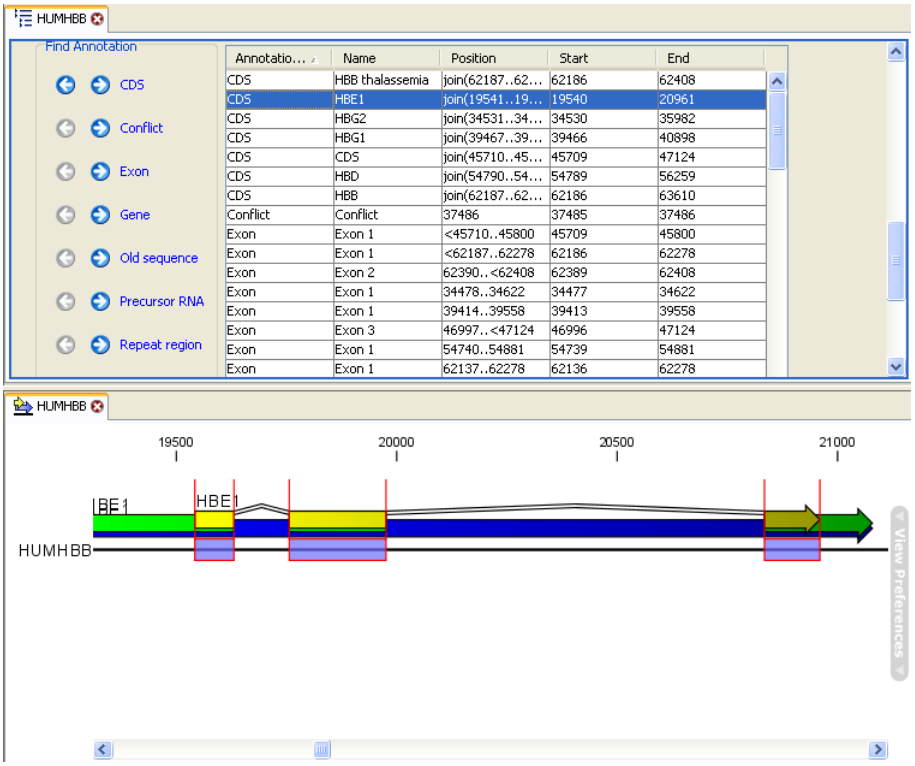


Figure 2.17: *Two views of the HUMHBB sequence. The upper view shows the coding sequences (CDS), and the bottom view shows a selection corresponding to the CDS chosen in the upper view.*

information about this molecule you wish to query the Swiss-Prot database to find homologous proteins in humans *Homo sapiens*, using the **Basic Local Alignment Search Tool** (BLAST) algorithm.
Please note that your computer must be connected to the Internet to complete this tutorial.

Start out by:

> **select protein NP_058652 in the Navigation Area | Toolbox | BLAST Search( ) | BLAST Against NCBI Databases**

In **Step 1** you can choose which sequence to use as query sequence. Since you have already chosen the sequence it is displayed in the **Selected Elements** list.

Click **Next**.

In **Step 2** (figure 2.18), choose the default BLAST program: **BLASTp: Protein sequence against Protein database** and select the **Swiss-Prot** database in the **Database** drop down menu.

Click **Next**.

In the **Limit by Entrez query** in **Step 3**, choose **Homo sapiens[ORGN]** from the drop down menu to arrive at the search configuration seen in figure 2.19. Including this term limits the query to proteins of human origin.

Click **Finish** to accept the default parameter settings and begin the BLAST search.

The computer now contacts NCBI and places your query in the BLAST search queue. After a short while the result is received and opened in a new view.

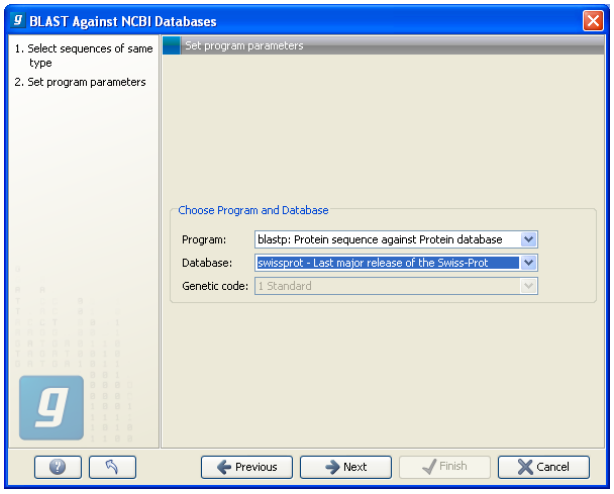


Figure 2.18: *Choosing BLAST program and database.*

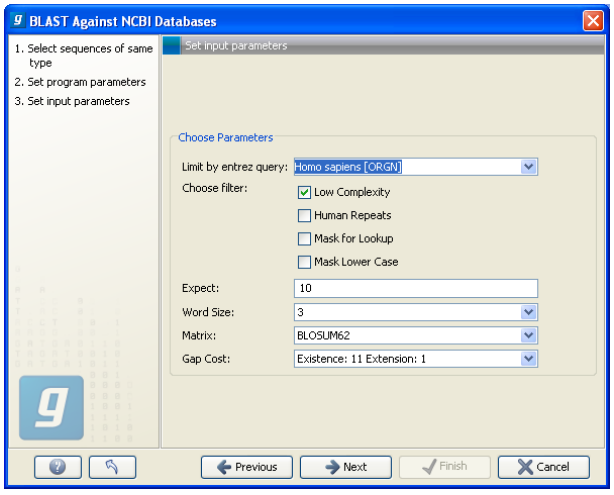


Figure 2.19: *The BLAST search is limited to homo sapiens[ORGN]. The remaining parameters are left as default.*

The output is shown in figure 2.20 and consists of a list of potential homologs that are sorted by their BLAST match-score and shown in descending order below the query sequence.

Try placing your mouse pointer over a potential homologous sequence. You will see that a context box appears containing information about the sequence and the match-scores obtained from the BLAST algorithm.

For now, we will focus our attention on sequence P02042 - the BLAST hit that is second from the top of the list. To open sequence P02042:

right-click the line representing sequence P02042 | Open Sequence in New View

This opens the sequence. However, the sequence is not saved yet. Drag and drop the sequence into the **Navigation Area** to save it. This homologous sequence is now part of your project and you can use it to gain information about the query sequence by using the various tools of the workbench, e.g. by studying its textual information, by studying its annotation or by aligning it to the query sequences.

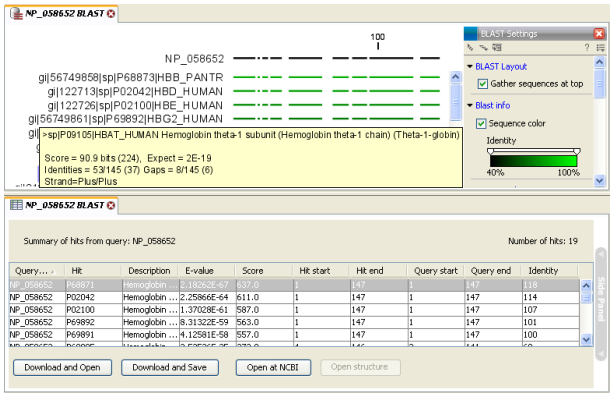

NP_058652 BLAST
100
NP_058652
gi|56749858|sp|P68873|HBB_PANTR
gi|122713|sp|P02042|HBD_HUMAN
gi|122726|sp|P02100|HBE_HUMAN
gi|56749861|sp|P69892|HBG2_HUMAN
>sp|P09105|HBAT_HUMAN Hemoglobin theta-1 subunit (Hemoglobin theta-1 chain) (Theta-1-globin)
Score = 90.9 bits (224), Expect = 2E-19
Identities = 53/145 (37) Gaps = 8/145 (6)
Strand=Plus/Plus
BLAST Settings
BLAST Layout
Gather sequences at top
Blast info
Sequence color
Identity
40%
100%
NP_058652 BLAST
Summary of hits from query: NP_058652
Number of hits: 19
Query... Hit Description E-value Score Hit start Hit end Query start Query end Identity
NP_058652 P68871 Hemoglobin ... 2.18262E-67 637.0 1 147 1 147 118
NP_058652 P02042 Hemoglobin ... 2.25866E-64 611.0 1 147 1 147 114
NP_058652 P02100 Hemoglobin ... 1.37028E-61 587.0 1 147 1 147 107
NP_058652 P69892 Hemoglobin ... 8.31322E-59 563.0 1 147 1 147 101
NP_058652 P69891 Hemoglobin ... 4.12581E-58 557.0 1 147 1 147 100
Download and Open
Download and Save
Open at NCBI
Open structure

Figure 2.20: *Output of a BLAST search. By holding the mouse pointer over the lines you can get information about the sequence.*

2.9 Tutorial: Primer design

In this tutorial, you will see how to use *CLC Gene Workbench* for finding primers for PCR amplification of a specific region.

The pBR322 sequence from the Example data is used in this tutorial. On positions 1891..1892 and 1913..1914 there are two "conflict" annotations which mark the region that should be amplified.

First, open the sequence in the Primer Designer:

Select the pBR322 sequence in the "Cloning Project" folder under "Nucleotide" in the Example data | Show( ) | Primer Designer( )

2.9.1 Finding the region to amplify

Next, make sure that the **conflict** annotations are shown (see figure 2.21):

Click Annotation types in the Side Panel | Conflict

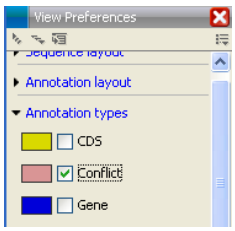


Figure 2.21: *The annotations of the type "conflict" are now visible.*

Find the annotation using the **Search** function in the **Side Panel**(see figure 2.22):

Search in the Side Panel | Type "conflict" in the text field | Check Annotation search | Click Search two times in order to go to position 1891 (we are not interested in the first conflict in this tutorial)

2.9.2 Specifying a region for the forward primer

The forward primer should be located in the region upstream from the annotation. Thus, select a region of approximately 40 residues, and mark it "Forward primer region here" (see figure 2.23):

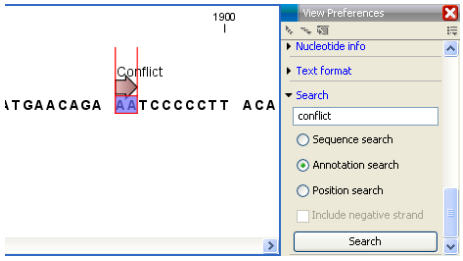


Figure 2.22: *The annotations of the type "conflict" are now visible.*

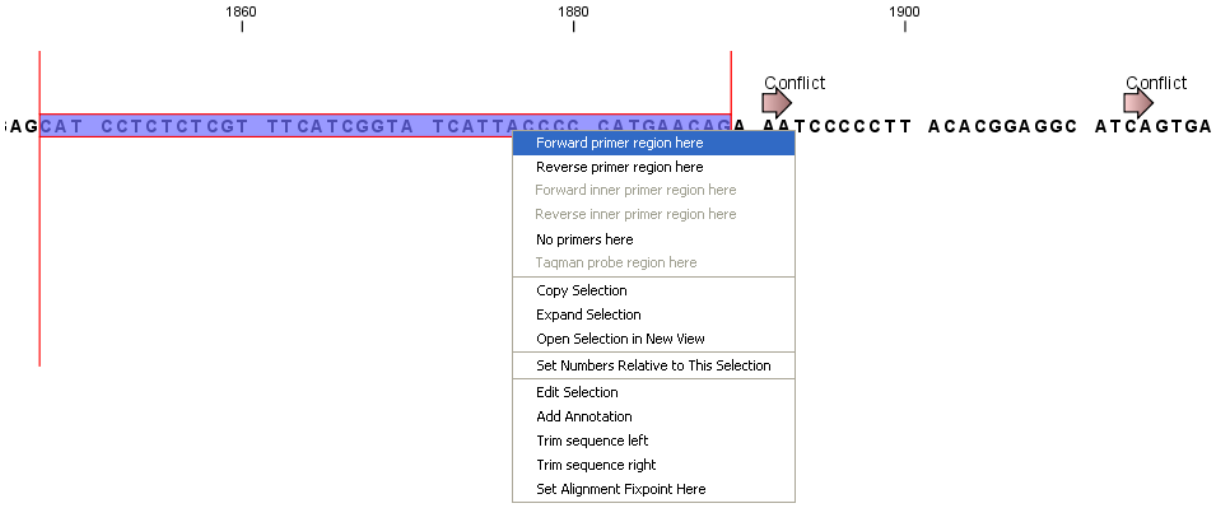


Figure 2.23: *Right-clicking a selection and choosing "Forward primer region here".*

2.9.3 Examining the primer suggestions

When you have specified this region, you will be able to see five lines of possible primers, based on the **Primer parameters** to the left. Each line represents primers of a specific length, e.g. the first line represents primers of length 18 (see figure 2.24):

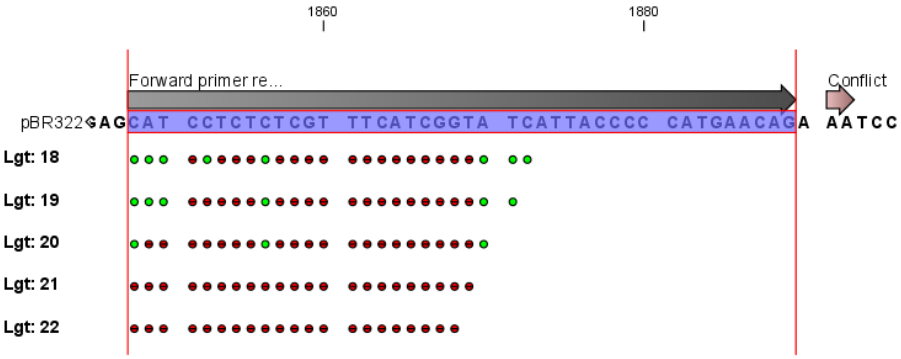


Figure 2.24: *Five lines of dots representing primer suggestions. There is a line for each length.*

Each line consists of a number of dots, each representing a possible primer. E.g. the first dot on the first line (primers of length 18) represents a primer starting at the dot's position and with a length of 18 nucleotides (shown as the white area in figure 2.25):

Position the mouse cursor upon a dot and you will see an information box providing data about this primer. Clicking the dot will select the region where the primer will anneal. (See figure 2.26):

Note that the dot is colored red and that there is an asterisk (*) before the melting temperature.

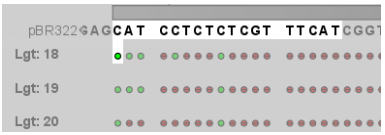


Figure 2.25: *The first dot on line one represents the starting point of a primer that will anneal to the highlighted region.*

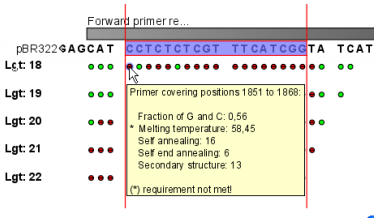


Figure 2.26: *Clicking the dot will select the corresponding region, and placing the cursor upon the dot will reveal an information box. .*

This indicates that the primer represented by this dot does not meet the requirements set in the **Primer parameters** (see figure 2.27):

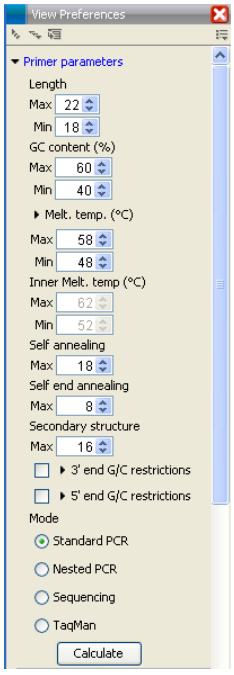


Figure 2.27: *The Primer parameters.*

Note that the maximum annealing temperature is per default set to 58, and this is the reason why the primer in figure 2.26 with an annealing temperature of 58,45 does not meet the requirements and is colored red. If you raise the maximum annealing temperature to 59, the primer will meet the requirements and the dot becomes green.

By adjusting the **Primer parameters** you can define primers which match your specific needs. Since the dots are constantly updated, you can immediately see how a change in the primer parameters affects the number of red and green dots.

2.9.4 Calculating a primer pair

Until now, we have been looking at the forward primer. To mark a region for the reverse primer, make a selection covering approximately 40 residues downstream of the conflict annotations, and:

Right-click the selection | Reverse primer region here

The two regions should now be located as shown in figure 2.28:

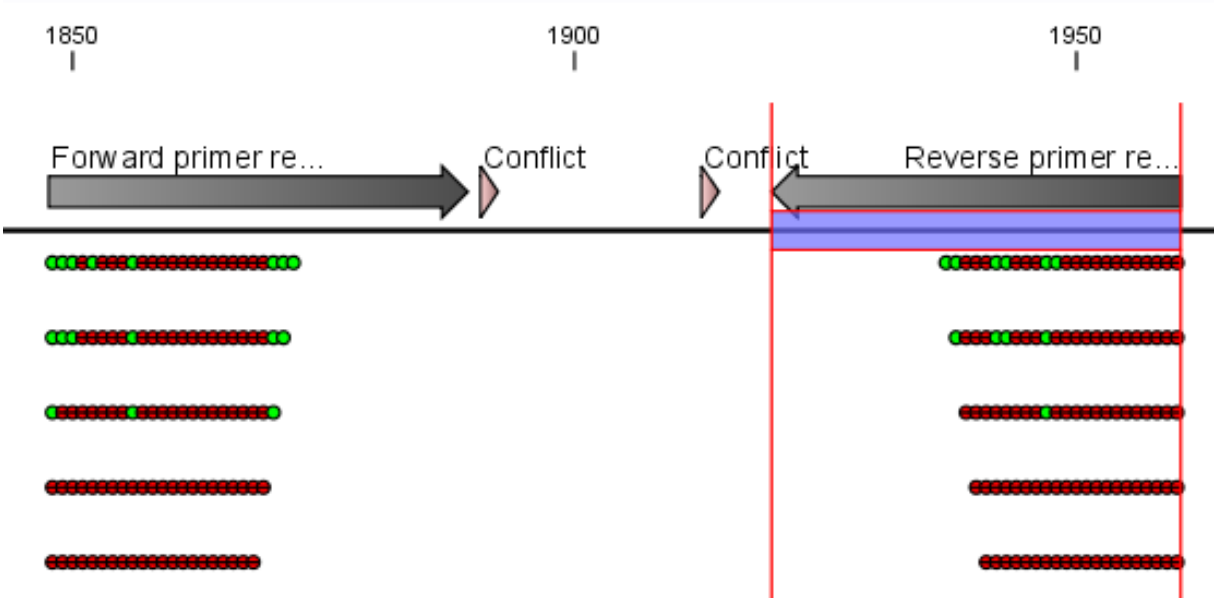


Figure 2.28: *A forward and a reverse primer region enclosing the conflicts.*

Now, you can let *CLC Gene Workbench* calculate all the possible primer pairs based on the **Primer parameters** that you have defined:

Click the Calculate button | Modify parameters regarding the combination of the primers (for now, just leave them unchanged)| Calculate

This will open a table showing the possible combinations of primers. To the right, you can specify the information you want to display, e.g. showing secondary structure (see figure 2.29):

Clicking a primer pair in the table will make a corresponding selection on the sequence in the view above. At this point, you can either settle on a specific primer pair or save the table for later. If you want to use e.g. the first primer pair for your experiment, right-click this primer pair in the table and save the primers.

You can also mark the position of the primers on the sequence by selecting **Mark primer annotation on sequence** in the right-click menu (see figure 2.30):

You have now reached the end of this tutorial which has shown some of the many options of the primer design functionalities of *CLC Gene Workbench*. You can read much more in the program's **Help** function ( ) or in the users manual on `http://www.clcbio.com/download`.

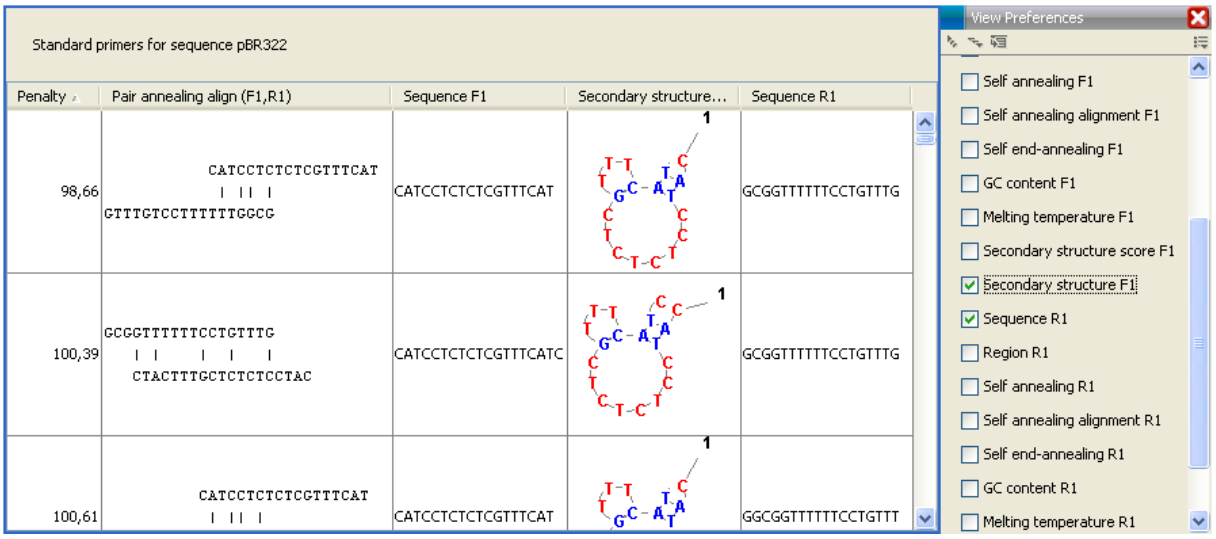


Figure 2.29: *A list of primers. To the right are the Side Panel showing the available choices of information to display.*

2.10 Tutorial: Assembly

In this tutorial, you will see how to assemble data from automated sequencers into a contig and how to find and inspect any inconsistencies that may exist between different reads.

First, select the five trace files (the reads) in the "Assembly" folder in the "Nucleotide" folder of the **Example data**. To assemble the files:

Toolbox in the Menu Bar | **Assembly (**) | **Assemble Sequences(**)

Click **Next** to go to the second step of the assembly where you choose to trim the sequences. In the next step, you will be able to specify how this trimming should be performed (see figure 2.31).

Leave these settings at their default and click **Finish**.

2.10.1 Getting an overview of the contig

The result of the assembly is a **Contig** which is an alignment of the five reads. Click **Fit width**( ) to see an overview of the contig. To help you determine the coverage, display a coverage graph (see figure 2.32):

Alignment info in Side Panel | **Coverage** | **Graph**

This overview can be an aid in determining whether coverage is satisfactory, and if not, which regions a new sequencing effort should focus on. Next, we go into the details of the contig.

2.10.2 Finding and editing inconsistencies

Click **Zoom to 100%**( ) to zoom in on the residues at the beginning of the contig. Click the **Find Inconsistency** button at the top of the **Side Panel** or press the **Space** key to find the first position where there is disagreement between the reads (see figure 2.33).

In this example, the first and the third reads have a "T", whereas the second line has a "C"

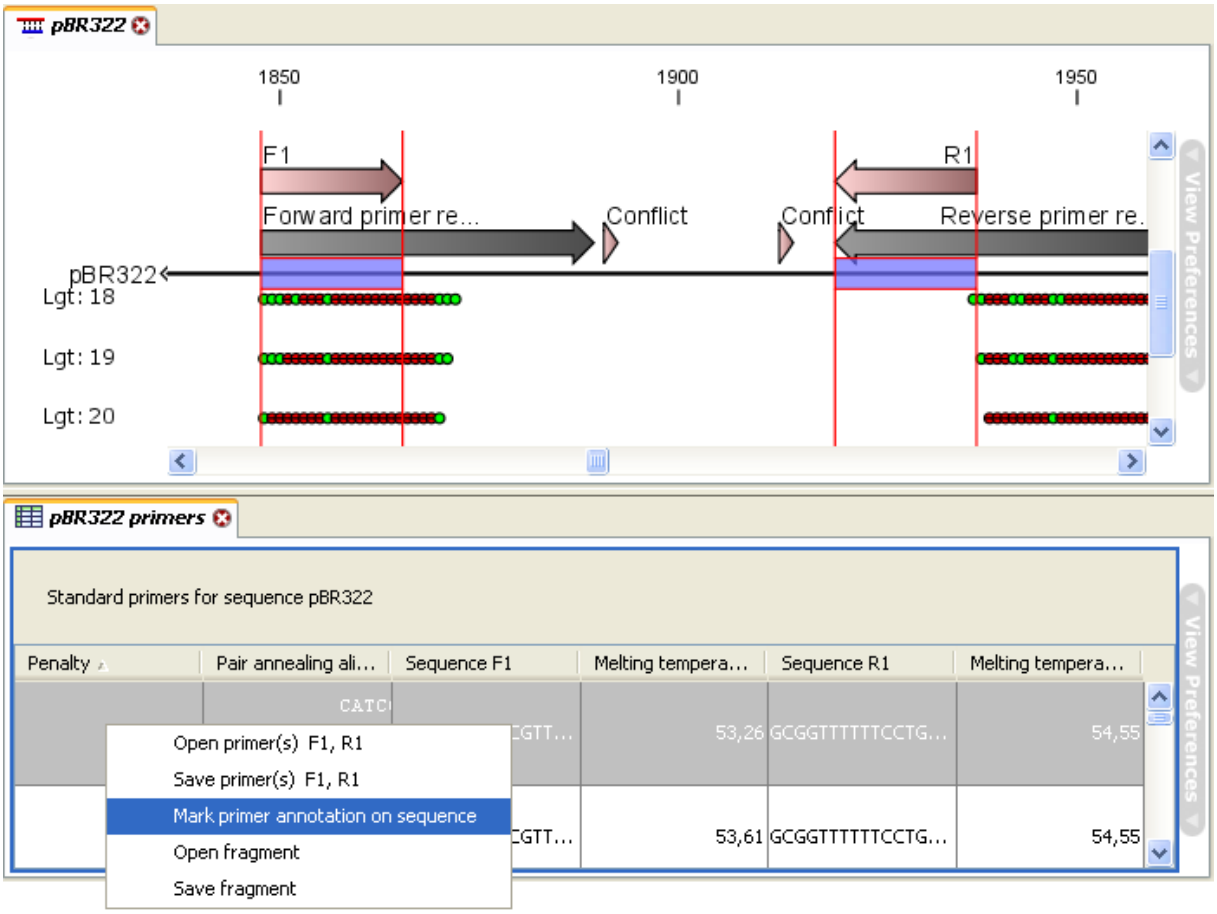


Figure 2.30: *The options available in the right-click menu. Here, "Mark primer annotation on sequence" has been chosen, resulting in two annotations on the sequence above (labeled F1 and R1).*

(marked with a light-pink background color). The gray color of the residues in the fourth line indicates that this region has been trimmed (based on the criteria in figure 2.31) and that this information is not included in the creation of the contig. Since the majority of the reads show a "T" in this position, we settle on this in the consensus. In order to show that there has been a disagreement in this position, type a lower-case "t" (see figure 2.34).

Clicking the **Find Inconsistency** button again will find the next inconsistency.

2.10.3 Documenting your changes

Whenever you make a change like replacing a "T" for a "t", it will be noted in the contig's history. To open the history:

Right-click the tab of the contig | Show | History( )

In the history, you can see the details of each change (see figure 2.35).

When you have finished editing the contig, the consensus can be saved:

Right-click the label "Contig" | Open Sequence in New View | Save( )

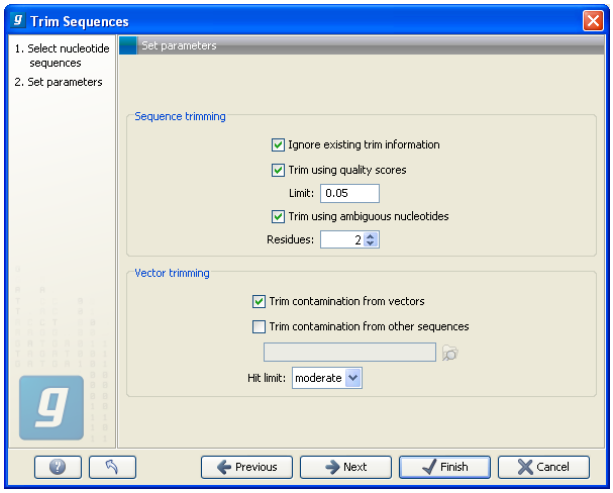


Figure 2.31: *Specifying how sequences should be trimmed.*

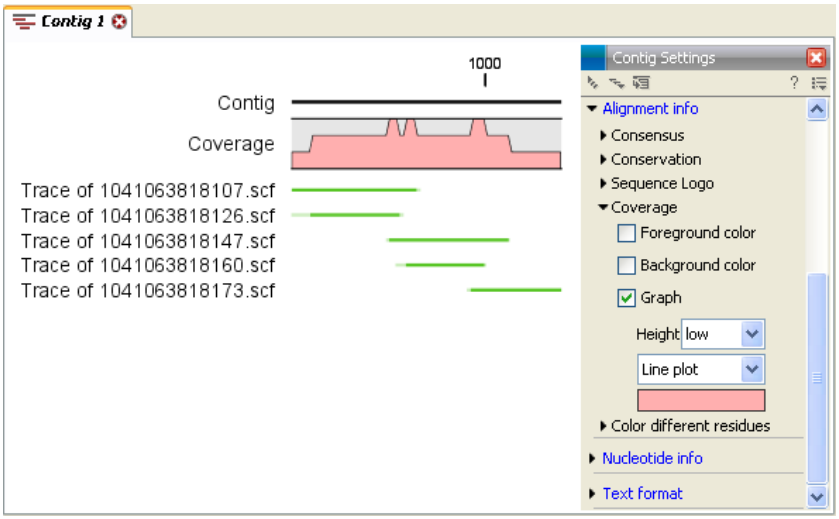


Figure 2.32: *An overview of the contig with the coverage graph.*

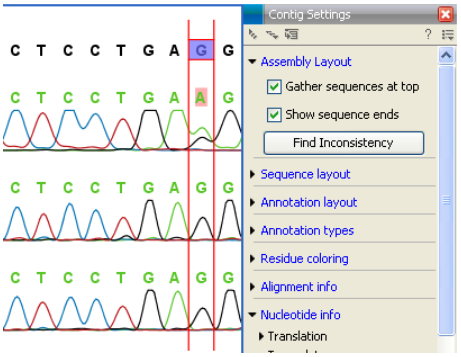


Figure 2.33: *Using the Find Inconsistency button highlights inconsistencies.*

2.11 Tips and tricks for the experienced user

In this tutorial you will get to know a number of ways to cut corners when using *CLC Gene Workbench*. The following sections will show you how to get your tasks done quickly and easily. When you are using the program it is hard to discover these shortcuts yourself which is the reason why this tutorial was written.

Figure 2.34: *Just press the key to replace the residue.*

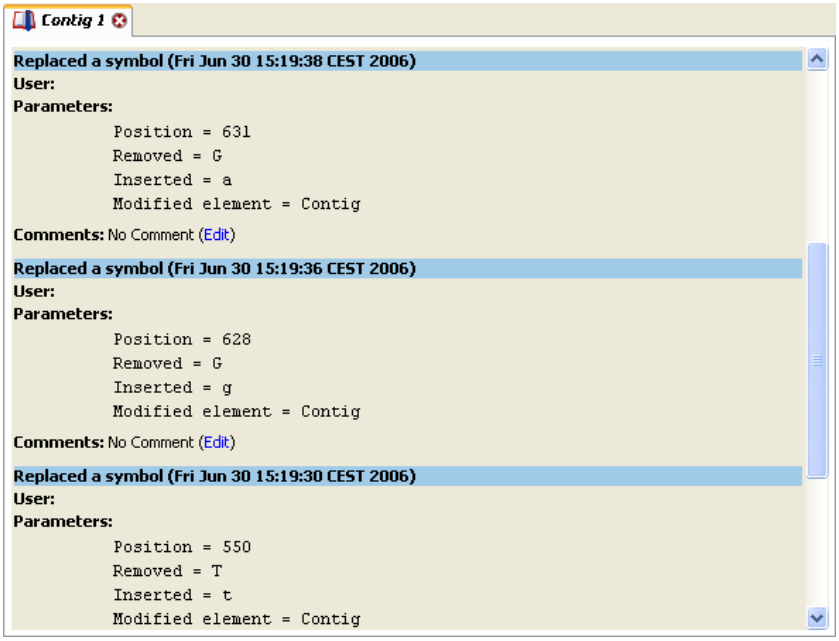


Figure 2.35: *The history of the contig showing that a "T" has been substituted for a "t" at position 550.*

The tutorial assumes that you have used the program for a while, since the basic usages are not explained.

2.11.1 Open and arrange views using drag and drop

Instead of opening views using double click or **Show**, you can use drag and drop both to open and arrange views. Drag and drop is supported both within the **Navigation Area**, within the **View Area** and between the two areas:

1. **Drag and drop an element within the Navigation Area:** Moves the element to the drop location.
2. **Drag an element from the Navigation Area to the View Area:** Opens the element in a new view. The view will be opened in the part of the **View Area** where the element is dropped.
3. **Drag the tab of a view within the View Area:** If there are other views open, this will split the **View Area** and make it possible to see several views at the time.
4. **Drag the tab of a view into the Navigation Area:** If the view is new and has not been saved to a project before, this will save the view at the drop location. If the view is already represented in the **Navigation Area**, this will save a copy of the view at the drop location.

2.11.2 Find element in the Navigation Area

If you have a view of e.g. a sequence and you wish to know in which project this sequence is saved, use the **Find in Project** function:

right-click the tab of the view | View | Find in Project( )

This will select the sequence in the **Navigation Area** (see figure 2.36).

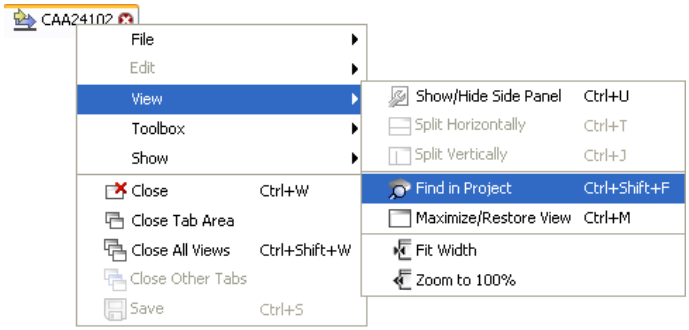


Figure 2.36: *This will select the sequence in the Navigation Area.*

You can also use the shortcut key: Ctrl + Shift + F on Windows or ⌘ + Shift + F on Mac.

2.11.3 Find specific annotations on a sequence

If you are looking for a specific annotation on a sequence, you may benefit from viewing the **Sequence info** while keeping an ordinary view of the sequence on the screen. In the **Sequence info** you find an Annotation map which displays all the annotations of the sequence. The annotations serve as links, selecting the annotation in the ordinary view of the sequence (see figure 2.37).

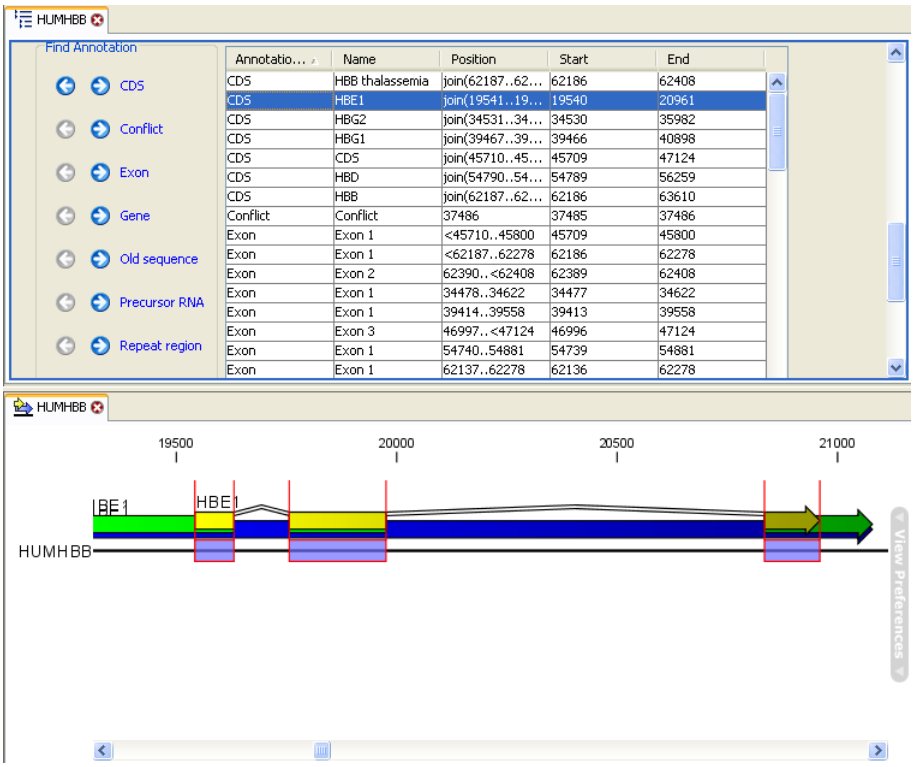


Figure 2.37: *Clicking the HBE1 coding region in the top view selects the annotation on the sequence in the bottom view.*

For sequences with many annotations, it is easier to navigate using these links compared to of scrolling in the ordinary view of the sequence.

2.11.4 Split sequences into several lines

Producing graphics of long sequences can be a strenuous task, especially if you have not discovered the "Wrap sequence" option. If you just export graphics of a long sequence without wrapping, you will get an extremely wide graphics file which probably has be edited in a graphics program before use. Wrapping the sequence allows you to control the width and height of the graphics file (see figure 2.38).

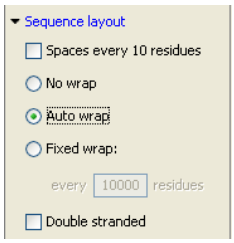


Figure 2.38: *Wrapping the sequence automatically.*

2.11.5 Make a new sequence of a coding region

If you have a genomic sequence containing a coding region, you can easily make a new sequence which only consists of the coding region (see figure 2.39):

right-click the coding region's annotation | Open Annotation in New View

This will open a new sequence which only consists of the residues covered by the annotation.

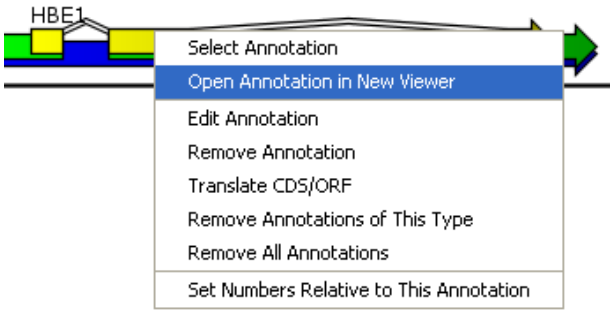


Figure 2.39: *Opening the coding region in a new view.*

2.11.6 Translate a coding region

If you have a genomic sequence containing one or more coding regions, you can translate these regions in a quick an easy way. If you want to translate a single coding region (see figure2.40):

right-click the coding region's annotation | Translate CDS/ORF

This will open a new view with the translated sequence.

In order to translate all the coding regions of a sequence:

Toolbox | Nucleotide Analyses ( ) | Translate to protein ( ) | Translate CDS and ORF in Step 2

This will extract all the coding regions of the sequence and for each region it will open a new view with the translation.

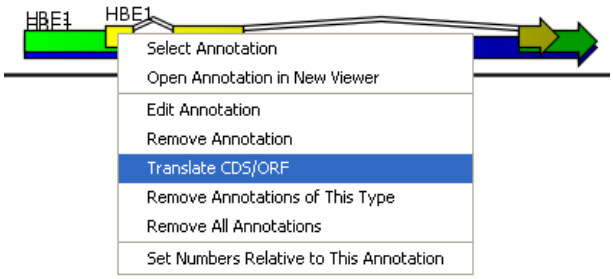


Figure 2.40: *Opening a new view with the translation of the coding region.*

2.11.7 Copy annotations from one sequence to another

If you have a collection of similar sequences, and you have annotated one of the sequences, you can copy these annotations to the rest of the sequences. First, create an alignment of the sequences. Next, find the annotated sequence and for each of the annotations that you want to copy:

right-click the annotation | Copy Annotation to other Sequences

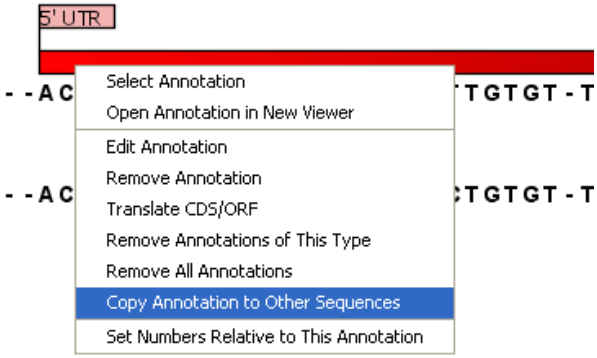


Figure 2.41: *Copying annotation to other sequences in the alignment.*

A dialog listing all the sequences in the alignment is shown. The annotation will be copied to the sequences that you select in this dialog. If the sequences are not identical, the annotation will still be copied.

2.11.8 Get overview and detail of a sequence at the same time

If you have a large sequence and you want to be able to get an overview of the whole and still keep the details of the residues, you can use the **Split views** functionality. In the example below (figure 2.42), the end of the red annotation is examined in detail in the bottom view, and in the upper view you have the overview of the whole alignment.

In this example, a selection was made in the upper view, and the bottom view automatically scrolls to display this selection (this behavior can be turned off by unchecking the "Follow selection" option in the **Side Panel**).

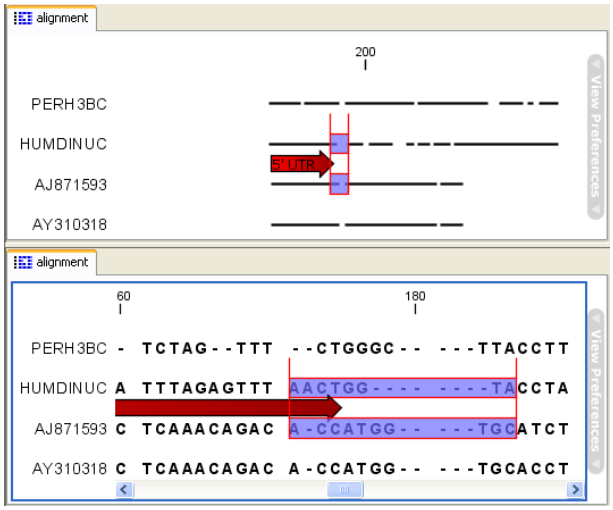


Figure 2.42: *Using the split views and follow selection functionalities.*

2.11.9 Smart selecting in sequences and alignments

There are a number of ways to select residues in sequences and alignments:

Using the mouse. This is the most basic way of selecting. Place the mouse cursor where you want the selection to start, press and hold the mouse button, move the mouse to the location where the selection should end and release the mouse button.

Using the mouse in combination with the Shift key. If you have made a selection and want to extend or reduce the selection, hold the Shift key while clicking the location where you want the boundary of the selection.

Using the arrow keys in combination with the Shift key. If you have made a selection and want to extend or reduce the selection, hold the Shift key while pressing the left and right arrow keys.

Using the mouse in combination with the Ctrl (for Windows) or ⌘ (for Mac) key. By holding this key, you can make multiple selections that are not contiguous.

Selecting an annotation. Double-click an annotation in order to select the residues that the annotation covers. This is especially helpful if the annotation is not contiguous (as the CDS region in figure 2.39).

Using the Search function. At the bottom of **Side Panel** to the right, there is a search field, which can be used for selections (use Ctrl + F on Windows or ⌘ + F on Mac). You can both search for annotations, residues or positions. The result of the search is a selection (as shown in figure 2.43). Remember to separate the start and end numbers with two punctuation marks (..).

No matter how you make your selection, you can see the start and end positions in right part of the status bar below the **View Area**.

2.11.10 Check for updates and additional information about sequences

If you have downloaded a sequence from NCBI or UniProt, you can easily check if the online information about the sequence has been updated and get additional information about the

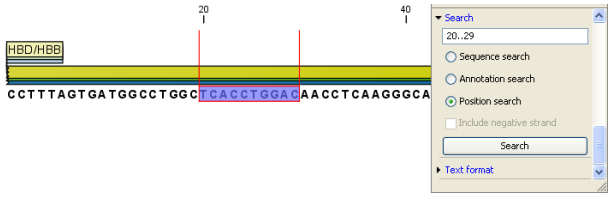


Figure 2.43: *Making a selection from position 20 to 29 (both included) using the Search function.*

sequence:

right-click the sequence | Web info | NCBI or UniProt

This will open your default web-browser showing the information about the sequence at either NCBI or UniProt. Clicking "PubMed" instead of "NCBI/UniProt" gives you a direct link to the sequence's PubMed references.

2.11.11 Quickly import sequences using copy-paste

Instead of using the **Import** ( ) function to import a sequence, you can use copy-paste. If you have copied the sequence from a source outside the program (e.g. a webpage or text document), you can paste it into the text field in the **Create new sequence** dialog (shown in figure 2.44).

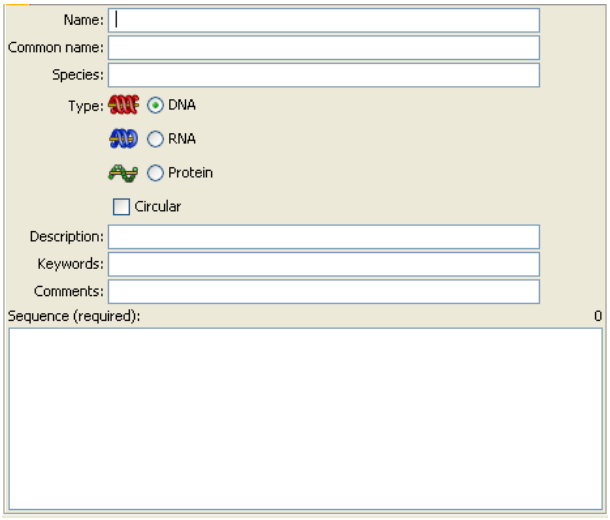


Figure 2.44: *Pasting a sequence into the text field at the bottom is a quick way of importing sequence data.*

This dialog lets you paste all kinds of characters into the text field, including numbers and spaces. If you have pasted e.g. numbers into the field, just press and hold the space key on your keyboard until the numbers have been deleted. Spaces are not included in the new sequence.

2.11.12 Perform analyses on many elements

If you have a folder with a lot of mixed elements (e.g. both nucleotide and protein sequences, alignments, reports), you can often select the whole folder for an analysis, even if the analysis should only be performed on a special type of element (e.g. nucleotide sequences). In the example below (figure 2.45), the dialog says "Select nucleotide sequences", but the project contains both protein and nucleotide sequences. Instead of carefully pinpointing the nucleotide

sequences, you can just press Ctrl+A (⌘ +A on Mac), selecting all the visible elements. When you add these elements (), the protein sequences are filtered out.

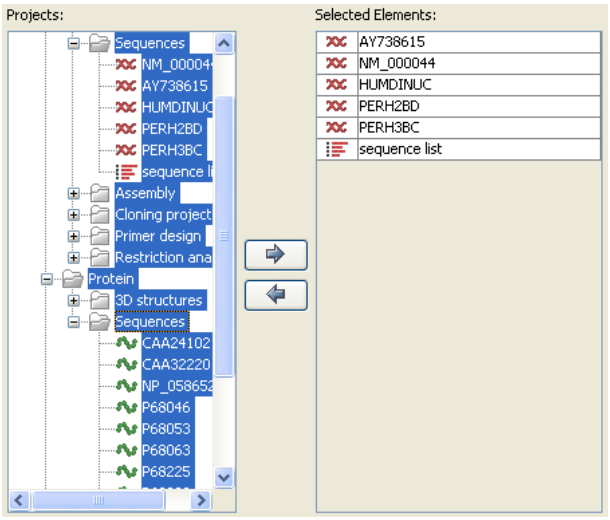


Figure 2.45: *Selecting protein and dna sequences, but the dialog automatically filters out the protein sequences.*

2.11.13 Drag elements to the Toolbox

If you have selected e.g. some protein sequences in the **Navigation Area** that you wish to use for creating an alignment,

2.11.14 Export elements while preserving history

If you have created e.g. an alignment and wish to export it to a colleague with the detailed history of all the source sequences, you can select the alignment and all the sequences for export. There is, however, a much easier way to do this (see figure 2.46):

Select the alignment | File | Export with dependent elements

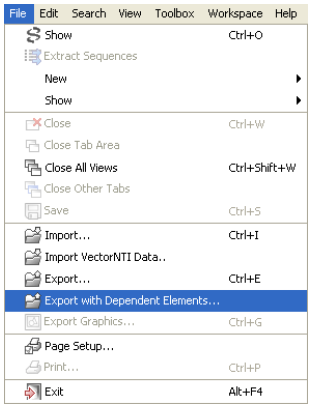


Figure 2.46: *Export with dependent elements in order to preserve the detailed history of an element.*

This will export the alignment including all the source sequences in one clc-file. When your colleague import the alignment, its detailed history is preserved.

2.11.15 Avoid the mouse trap - use keyboard shortcuts

Many tasks can be performed without using the mouse. When you do the same task again and again, you can save some time by learning its shortcut key. As an example you can navigate and zoom a view of sequence or an alignment using the keyboard:

- **Navigate the view using the four arrow keys.** This is equivalent to scrolling with the mouse using the scroll bars.
- **Use the '+' and '-' keys to zoom in and out.** This is equivalent to using the zoom modes in the toolbar.

Note that you have to click once inside the view with the mouse first in order to use this functionality.

There are many other shortcuts in *CLC Gene Workbench*which may save you a lot of time when performing repetitive tasks. See section 3.6 for a list of available shortcuts.

Part II

Basic Program Functionalities

Chapter 3

User Interface

Contents

3.1 Navigation Area 58
3.1.1 Data structure 58
3.1.2 Create new projects and folders 59
3.1.3 Multiselecting elements 60
3.1.4 Moving and copying elements 60
3.1.5 Change element names 61
3.1.6 Delete elements 62
3.1.7 Show folder elements in View 63
3.1.8 Sequence properties 64
3.2 View Area 64
3.2.1 Open View 64
3.2.2 Close Views 65
3.2.3 Save changes in a View 66
3.2.4 Undo/Redo 66
3.2.5 Arrange Views in View Area 67
3.2.6 Side Panel 68
3.3 Zoom and selection in View Area 69
3.3.1 Zoom In 69
3.3.2 Zoom Out 71
3.3.3 Fit Width 71
3.3.4 Zoom to 100% 71
3.3.5 Move 71
3.3.6 Selection 71
3.4 Toolbox and Status Bar 72
3.4.1 Processes 72
3.4.2 Toolbox 72
3.4.3 Status Bar 73
3.5 Workspace 73
3.5.1 Create Workspace 73
3.5.2 Select Workspace 73

3.5.3 Delete Workspace 73
3.6 List of shortcuts 74

This chapter provides an overview of the different areas in the user interface of *CLC Gene Workbench 2.0*. As can be seen from figure 3.1 this includes a **Navigation Area**, **View Area**, **Menu Bar**, **Toolbar**, **Status Bar** and **Toolbox**.

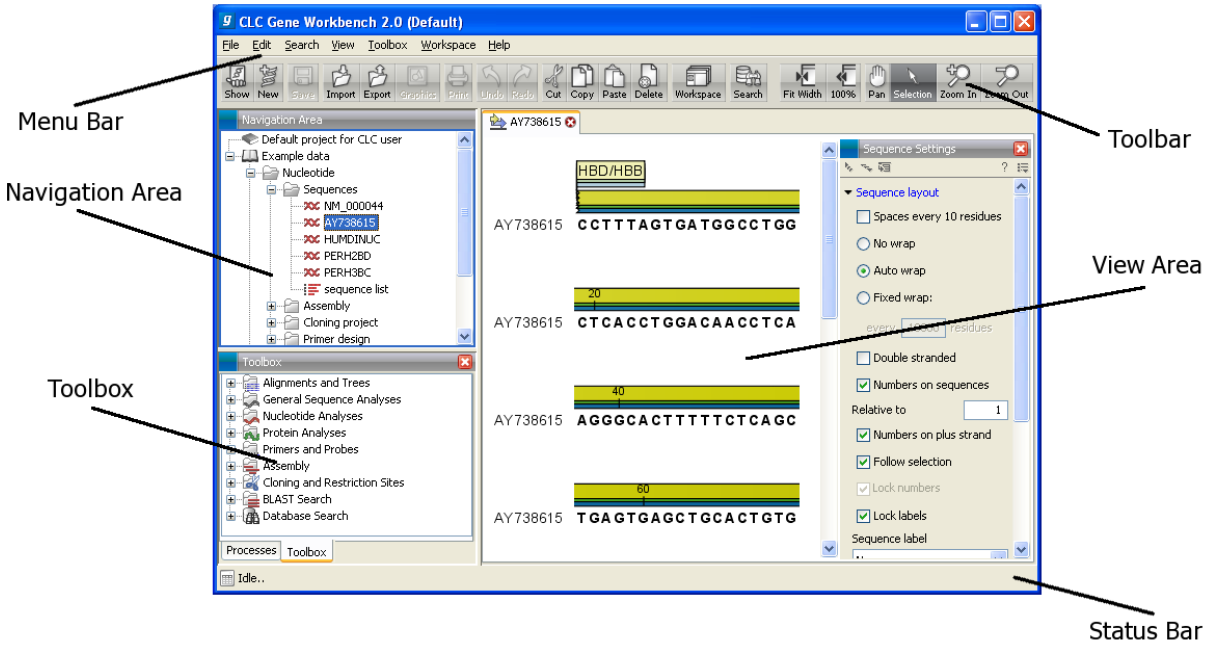


Figure 3.1: *The user interface consists of the Menu Bar, Toolbar, Status Bar, Navigation Area, Toolbox, and View Area.*

3.1 Navigation Area

The **Navigation Area** is located in the left side of the workbench, under the **Toolbar**. It is used for organizing and navigating data. The **Navigation Area** displays a **Project Tree** (see figure 3.2), which is similar to the way files and folders are usually displayed on your computer. The **Project Tree** contains one or more projects. The elements which are available in the **Navigation Area** remain the same when changing **Workspaces** (see section 3.5).

A project can be a collection of elements which are related, e.g. because the elements are used in the same assignment or research project.

The word 'Element' is used to refer to sequences, saved searches, lists, folders etc. In other words, everything which can be stored in a project in the **Navigation Area**.

3.1.1 Data structure

Elements, or data, in *CLC Gene Workbench 2.0* are stored in a kind of database. Hence, the data cannot be browsed from e.g. Windows Explorer or similar file systems. However, elements are available from the **Navigation Area**. To open an element:

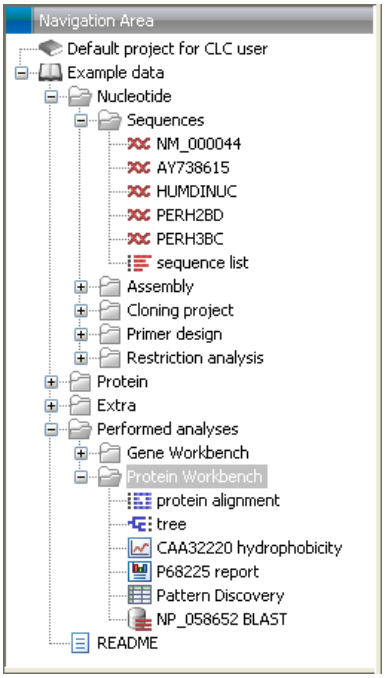


Figure 3.2: *The Navigation Area.*

Double-click the element

or **Click the element | Show ( ) in the Toolbar | Select the desired way to view the element**

This will open a **View** in the **View Area**, which is described in the next section.

Adding data

Data can be added to a project in a number of ways. Files can be imported from the file system, and elements from the **Navigation Area** can also be exported to the file system. (For more about import and export, see chapter 6.)

Furthermore, an element can be added to a project by dragging it into the **Navigation Area**. Elements on lists, e.g. search hits or sequence lists, can also be dragged to the **Navigation Area**.

When dragging from the **View Area** to the **Navigation Area**, the element, e.g. a sequence, an alignment, or a search report, is selected by clicking on the tab and dragging it into the navigation area. If the element already exists, you are asked whether you want to save a copy.

If a piece of data is dropped on a folder or a project, the data is placed at the bottom of the list of elements in the folder or project in question.

If a piece of data is dropped on an element, which is not a folder or a project, the data is added just after that element.

3.1.2 Create new projects and folders

In the **Navigation Area** all files and folders are stored in one or more projects. Creating a new project can be done in two ways:

right-click an element in the Navigation Area | **New** | **New Project (** **)**

or **File** | **New** | **New Project (** **)**

Regardless of which element is selected when you create a new project, the new project is placed at the bottom of the **Project Tree**.

You can move the project manually by selecting it and dragging it to the desired location. Projects are always placed at the upper-most level in the **Project Tree**.

In order to organize your files, they can be placed in folders. Creating a new folder can be done in two ways:

right-click an element in the Navigation Area | **New** | **New Folder (** **)**

or **File** | **New** | **New Folder (** **)**

If a project or a folder is selected in the **Navigation Area** when adding a new folder, the new folder is added at the bottom of the project or folder. If an element is selected, the new folder is added right below that element.

You can move the folder manually by selecting it and dragging it to the desired location.

3.1.3 Multiselecting elements

Multiselecting elements in the **Navigation Area** can be done in the following ways:

- Holding down the <Ctrl> key while clicking on multiple elements selects the elements that have been clicked.
- Selecting one element, and selecting another element while holding down the <Shift> key selects all the elements listed between the two locations (the two end locations included).
- Selecting one element, and moving the curser with the arrow-keys while holding down the <Shift> key, enables you to increase the number of elements selected.

3.1.4 Moving and copying elements

Elements can be moved and copied in two ways: using the copy, cut and paste functions, or using drag and drop.

Copy, cut and paste elements

Copies of elements, folders, and projects can be made with the copy/paste function which can be applied in a number of ways:

select the files to copy | **right-click one of the selected files** | **Copy (** **)** | **right-click the location to insert files into** | **Paste (** **)**

or **select the files to copy** | **Ctrl + C (⌘ + C on Mac)** | **select where to insert files** | **Ctrl + P (⌘ + P on Mac)**

or **select the files to copy** | **Edit in the Menu Bar** | **Copy (** **)** | **select where to insert files** | **Edit in the Menu Bar** | **Paste (** **)**

If there is already an element of that name, the pasted element will be renamed by appending a number at the end of the name. Elements can also be moved instead of copied. This is done with the cut/paste function.

select the files to cut | **right-click one of the selected files** | **Cut ( )** | **right-click the location to insert files into** | **Paste ( )**

or **select the files to cut** | **Ctrl + X (⌘ + X on Mac)** | **select where to insert files** | **Ctrl + V (⌘ + V on Mac)**

When you have cut the element, it disappears until you activate the paste function.

Move using drag and drop

Using drag and drop in the **Navigation Area**, as well as in general, is a four-step process:

click the element | **click on the element again, and hold left mouse button** | **drag the element to the desired location** | **let go of mouse button**

This allows you to:

- Move elements between different projects and folders in the **Project Tree**
- Drag from the **Navigation Area** to the **View Area**: A new **View** is opened in an existing **View Area** if the element is dragged from the **Navigation Area** and dropped next to the tab(s) in that **View Area**.
- Drag from the **View Area** to the **Navigation Area**: The element, e.g. a sequence, alignment, search report etc. is saved where it is dropped. If the element already exists, you are asked whether you want to save a copy. You drag from the **View Area** by dragging the tab of the desired element.

Use of drag and drop is supported throughout the program. Further description of the function is found in connection with the relevant functions.

3.1.5 Change element names

This section describes two ways of changing the names of sequences in the **Navigation Area**. In the first part, the sequences themselves are not changed - it's their representation that changes. The second part describes how to change the name of the element.

Change how sequences are displayed

Sequence elements can be displayed in the **Navigation Area** with different types of information:

- Name (this is the default information to be shown).
- Accession (sequences downloaded from databases like GenBank have an accession number).
- Species.

- Species (accession).
- Common Species.
- Common Species (accession).

Whether sequences can be displayed with this information depends on their origin. Sequences that you have created yourself or imported might not include this information, and you will only be able to see them represented by their name. However, sequences downloaded from databases like GenBank will include this information. To change how sequences are displayed:

right-click any element or folder in the Navigation Area | Sequence Representation | select format

This will only affect sequence elements, and the display of other types of elements, e.g. alignments, trees and external files, will be not be changed. If a sequence does not have this information, there will be no text next to the sequence icon.

Rename element

Renaming a project, folder, piece of data etc. can be done in three different ways:

right-click the element | Rename

or **select the element | Edit in the Menu Bar | Rename**

or **select the element | F2**

When the editing of the name has finished; press enter or select another element in the **Navigation Area**. If you want to discard the changes instead, press the **Esc**-key.

3.1.6 Delete elements

Deleting a project, folder, piece of data, etc. can be done in two ways:

right-click the element | Delete (**)**

or **select the element | press Delete key**

This will cause the element to be moved to a **Recycle Bin** where it is kept as a precaution.

Restore Deleted Elements

The elements in the **Recycle Bin** can be restored and saved in the **Navigation Area** again. This is done by:

Edit in the Menu Bar | Restore Deleted Elements (**)**

This opens the dialog shown in fig. 3.3.

The dialog shows a list of all the deleted elements. Select the elements you want to restore and click next. This opens the dialog shown in fig. 3.4.

Choose where to restore the deleted elements. Click **Finish**

Notice! Only files which were saved in the **Navigation Area**, and then deleted, can be restored.

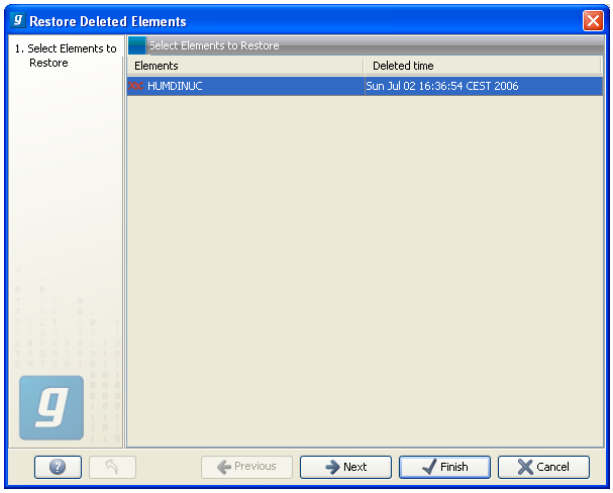


Figure 3.3: *The Restore Deleted Elements dialog.*

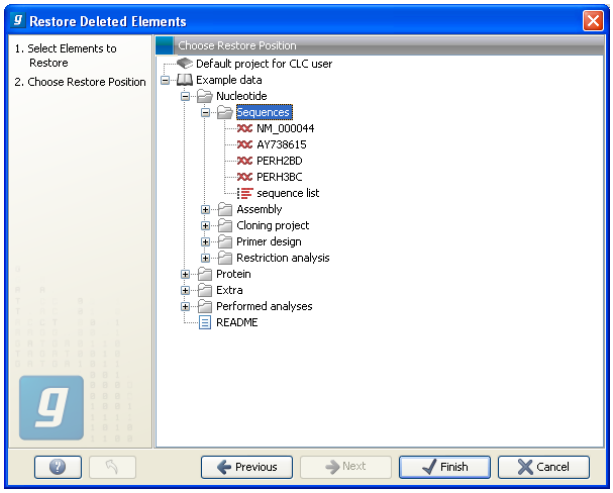


Figure 3.4: *The Restore Deleted Elements dialog.*

The deleted elements remain in the **Recycle Bin** until the **Recycle Bin** is emptied. To empty the bin:

Edit in the Menu Bar | **Empty recycle bin (** **)**

3.1.7 Show folder elements in View

A project or a folder might contain large amounts of elements. It is possible to view the elements of a folder or project in the **View Area**:

select a project | **Show (** **) in the Toolbar** | **Folder Contents (** **)**

When the elements are shown in the **View**, they can be sorted by clicking the heading of each of the columns. You can further refine the sorting by pressing Ctrl while clicking the heading of another column.

Sorting the elements in a **View** does not affect the ordering of the elements in the **Navigation Area**.

Notice! The **View** only displays one layer of the **Project Tree** at a time.

3.1.8 Sequence properties

Sequences downloaded from databases have a number of properties, which can be displayed using the **Sequence Properties** function:

Right-click a sequence in the Navigation Area |Properties

This will show a dialog as shown in figure 3.5.

Figure 3.5: *Sequence properties for the HUMDINUC sequence.*

For a more comprehensive view of sequence information, see section 11.2.

3.2 View Area

The **View Area** is the right-hand part of the workbench interface, displaying your current work. The **View Area** may consist of one or more **Views**, represented by tabs at the top of the **View Area**.

This is illustrated in figure 3.6.

Notice I.e., the tab concept is central to working with *CLC Gene Workbench 2.0*, because several operations can be performed by dragging the tab of a view, and extended right-click menus can be activated from the tabs.

This chapter deals with the handling of **Views** inside a **View Area**. Furthermore, it deals with rearranging the **Views**.

Section 3.3 deals with the zooming and selecting functions.

3.2.1 Open View

Opening a **View** can be done in a number of ways:

double-click an element in the Navigation Area

or **select an element in the Navigation Area | File | Show | Select the desired way to view the element**

or **select an element in the Navigation Area | Ctrl + O (⌘ + B on Mac)**

Opening a **View** while another **View** is already open, will show the new **View** in front of the other **View**. The **View** that was already open can be brought to front by clicking its tab.

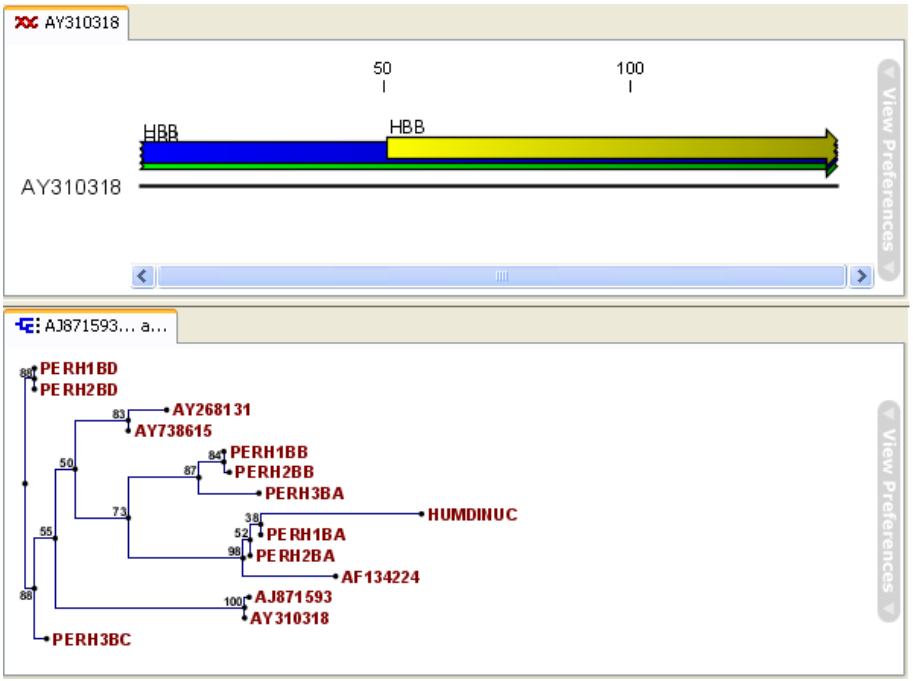


Figure 3.6: *A View Area can enclose several Views, each View is indicated with a tab (see top left View, which shows protein P12675). Furthermore, several Views can be shown at the same time (in this example, three views are displayed).*

Notice! If you right-click an open tab of any element, click **Show**, and then choose a different view of the same element, this new view is automatically opened in a split-view, allowing you to see both views.

See section 3.1.4 for instructions on how to open a **View** using drag and drop.

3.2.2 Close Views

When a **View** is closed, the **View Area** remains open as long as there is at least one open **View**.

A **View** is closed by:

right-click the tab of the View | Close

or **select the View | Ctrl + W**

or **hold down the Ctrl-button | Click the tab of the view while the button is pressed**

By right-clicking a tab, the following close options exist. See figure 3.7

- **Close.** See above.
- **Close Tab Area.** Closes all tabs in the tab area.
- **Close All Views.** Closes all tabs, in all tab areas. Leaves an empty workspace.
- **Close Other Tabs.** Closes all other tabs in the particular tab area.

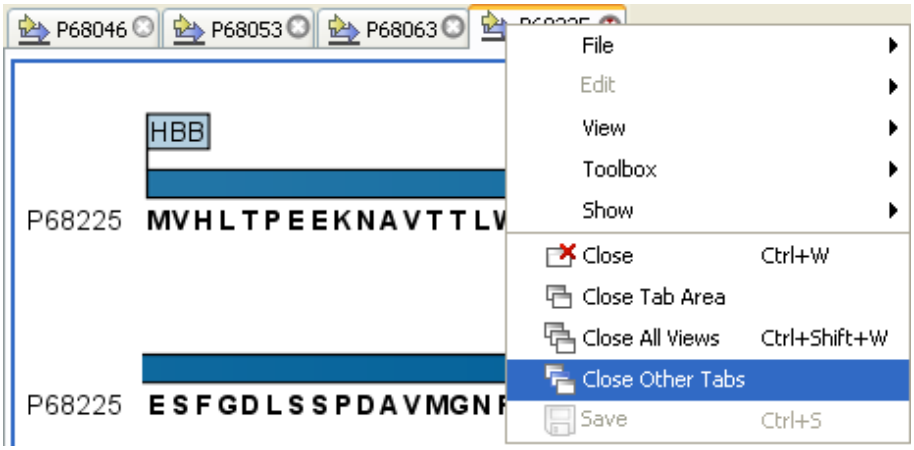


Figure 3.7: *By right-clicking a tab, several close options are available.*

3.2.3 Save changes in a View

When changes are made in a view, the text on the tab appears *bold and italic*. This indicates that the changes are not saved. The **Save** function may be activated in two ways:

Click the tab of the View you want to save | Save ( ) in the toolbar.

or **Click the tab of the View you want to save | Ctrl + S (⌘ + S on Mac)**

If you close a **View** containing an element that has been changed since you opened it, you are asked if you want to save.

When saving a new view that has not been opened from the Navigation Area (e.g. when opening a sequence from a list of search hits), a save dialog appears (figure 3.8).

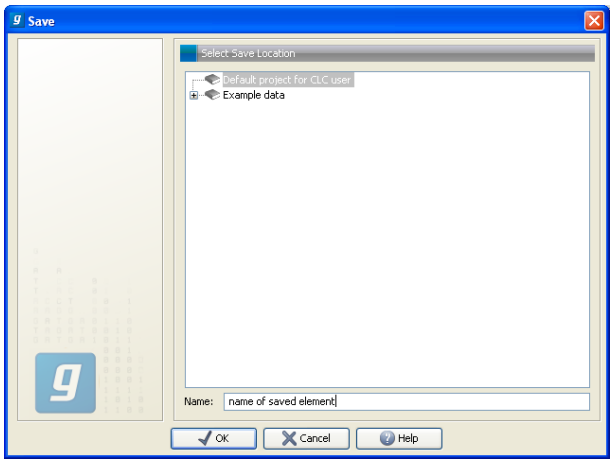


Figure 3.8: *Save dialog.*

In the dialog you select the folder or project in which you want to save the element.

After naming the element, press **OK**

3.2.4 Undo/Redo

If you make a change in a view, e.g. remove an annotation in a sequence or modify a tree, you can undo the action. In general, **Undo** applies to all changes you can make when right-clicking in

a view. **Undo** is done by:

Click undo ( ) in the Toolbar

or **Edit** | **Undo ( )**

or **Ctrl + Z**

If you want to undo several actions, just repeat the steps above. To reverse the undo action:

Click the redo icon in the Toolbar

or **Edit** | **Redo ( )**

or **Ctrl + Y**

Notice! Actions in the **Navigation Area**, e.g. renaming and moving elements, cannot be undone. However, you can restore deleted elements (see section 3.1.6).

You can set the number of possible undo actions in the Preferences dialog (see section 4).

3.2.5 Arrange Views in View Area

Views are arranged in the **View Area** by their tabs. The order of the **Views** can be changed using drag and drop. E.g. drag the tab of one **View** onto the tab of a another. The tab of the first **View** is now placed at the right side of the other tab.

If a tab is dragged into a **View**, an area of the **View** is made gray (see fig. 3.9) illustrating that the view will be placed in this part of the **View Area**.

The results of this action is illustrated in figure 3.10.

You can also split a **View Area** horizontally or vertically using the menus.

Splitting horisontally may be done this way:

right-click a tab of the View | **View** | **Split Horizontally ( )**

This action opens the chosen **View** below the existing **View**. (See figure 3.11). When the split is made vertically, the new **View** opens to the right of the existing **View**.

Splitting the **View Area** can be undone by dragging e.g. the tab of the bottom view to the tab of the top view. This is marked by a gray area on the top of the view.

Maximize/Restore size of View

The **Maximize/Restore View** function allows you to see a **View** in maximized mode, meaning a mode where no other **Views** nor the **Navigation Area** is shown.

Maximizing a **View** can be done in the following ways:

select View | **Ctrl + M**

or **select View** | **View** | **Maximize/restore size of View ( )**

or **select View** | **right-click the tab** | **View** | **Maximize/restore View ( )**

or **double-click the tab of View**

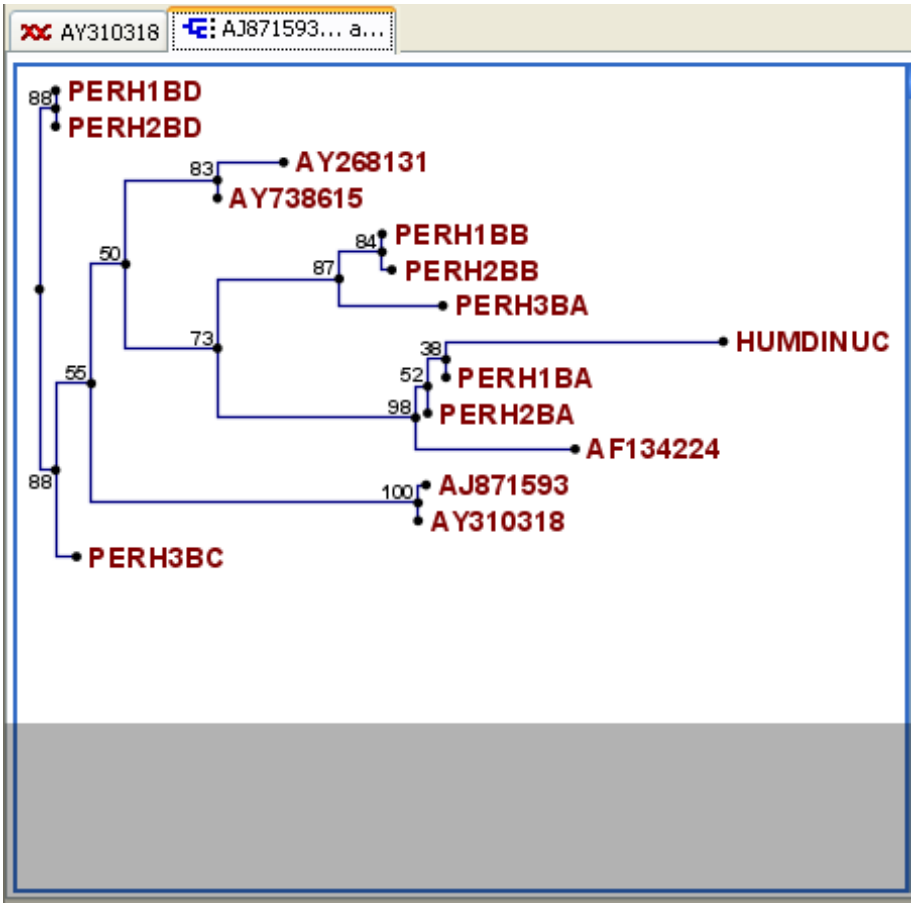


Figure 3.9: *When dragging a View, a gray area indicates where the View will be shown.*

The following restores the size of the **View**:

Ctrl + M

or **View** | **Maximize/restore size of View (** **)**

or **click close-button (** **) in the corner of the View Area**

or **double-click title of View**

3.2.6 Side Panel

The **Side Panel** allows you to change the way the contents of a view are displayed. The options in the **Side Panel** depend on the kind of data in the **View**, and they are described in the relevant sections about sequences, alignments, trees etc.

Side Panel are activated in this way:

select the View | **Ctrl + U (⌘ + U on Mac)**

or **right-click the tab of the View** | **View** | **Show/Hide Side Panel (** **)**

Notice! Changes made to the **Side Panel** will not be saved when you save the **View**. See how to save the changes in the **Side Panel** in chapter 4 .

The **Side Panel** consists of a number of groups of preferences (depending on the kind of data

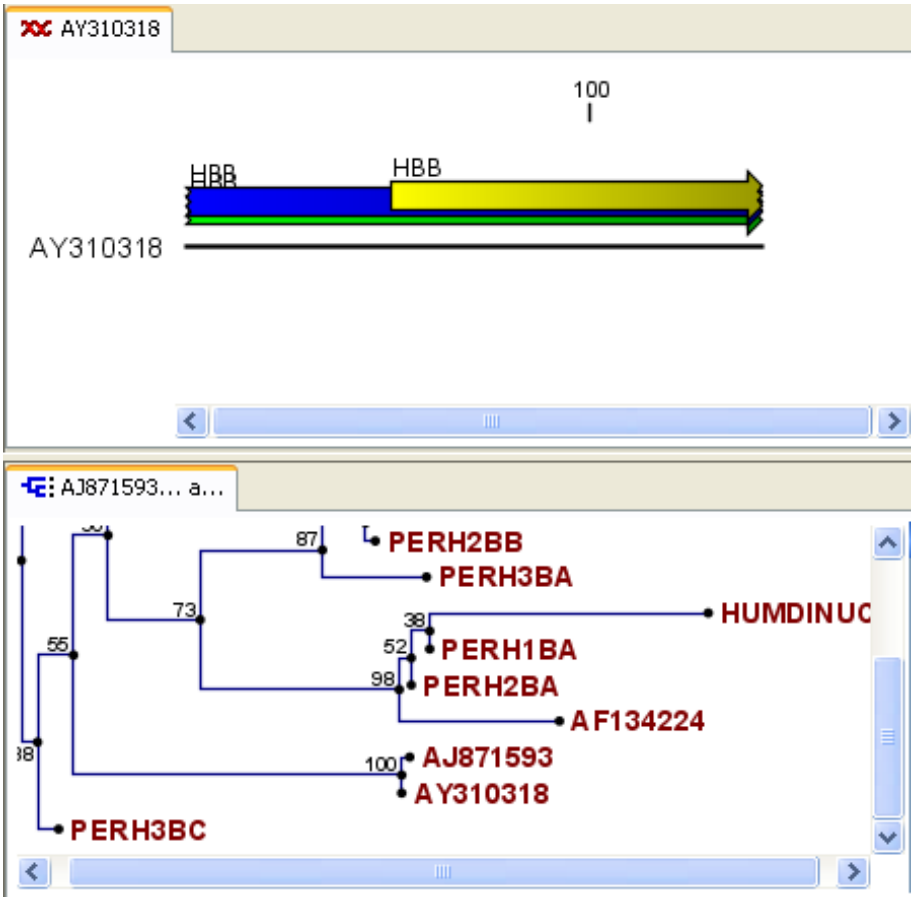


Figure 3.10: *A horizontal split-screen. The two Views split the View Area.*

being viewed), which can be expanded and collapsed by clicking the header of the group. You can also expand or collapse all the groups by clicking the icons ( )/( ) at the top.

3.3 Zoom and selection in View Area

The mode toolbar items in the right side of the **Toolbar** apply to the function of the mouse pointer. When e.g. **Zoom Out** is selected, the **Zoom Out**-function is applied each time you click in a **View** where zooming is relevant (texts, tables and lists cannot be zoomed). The chosen mode is active until another mode toolbar item is selected. (**Fit Width** and **Zoom to 100%** do not apply to the mouse pointer.)

3.3.1 Zoom In

There are two ways to **Zoom In**:

The first way enables you to zoom in, step by step, on a sequence:

Click Zoom In ( ) in the toolbar | click the location in the view that you want to zoom in on

or **Click Zoom In ( ) in the toolbar | click-and-drag a box around a part of the view | the view now zooms in on the part you selected**

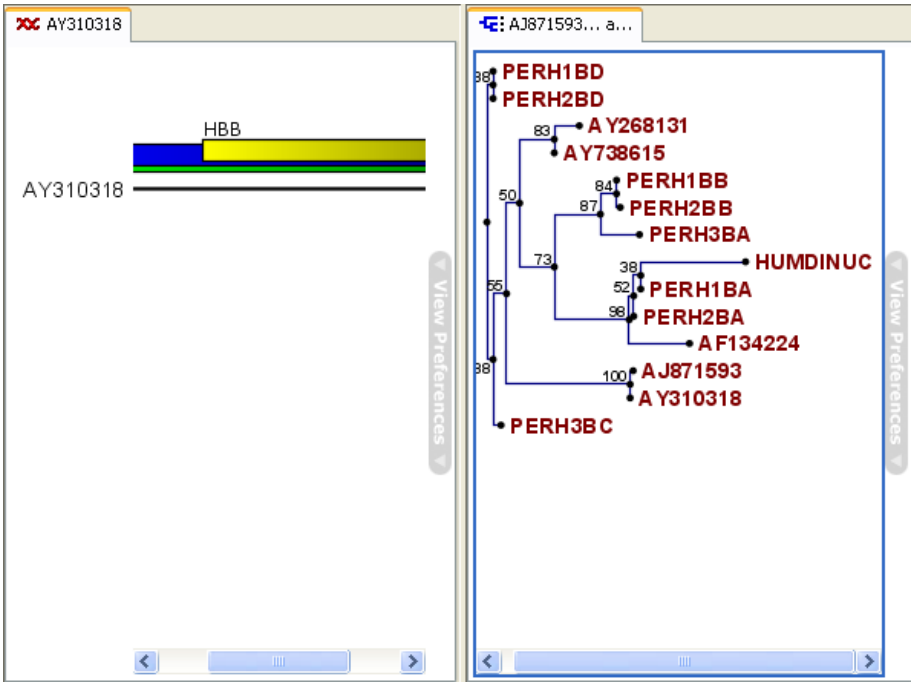


Figure 3.11: *A vertical split-screen.*

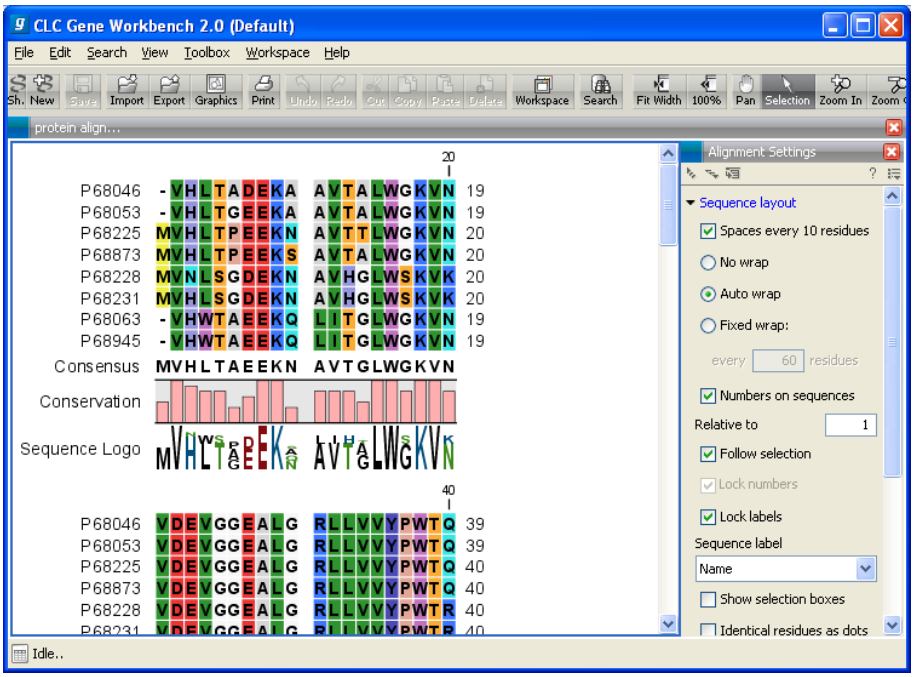


Figure 3.12: *A maximized View. The function hides the Navigation Area and the Toolbox.*

When you choose the Zoom In mode, the mouse pointer changes to a magnifying glass to reflect the mouse mode.

If you press the **Shift** button on your keyboard while clicking in a **View**, the zoom funtion is reversed. Hence, clicking on a sequence in this way while the **Zoom In** mode toolbar item is selected, zooms out instead of zooming in.

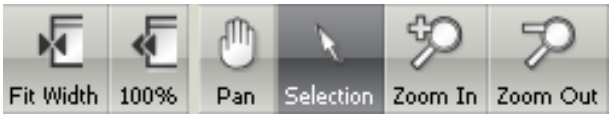


Figure 3.13: *The mode toolbar items.*

3.3.2 Zoom Out

It is possible to zoom out, step by step, on a sequence:

> **Click Zoom Out ( ) in the toolbar | click in the view until you reach a satisfying zoomlevel.**

When you choose the Zoom In mode, the mouse pointer changes to a magnifying glass to reflect the mouse mode.

If you want to get a quick overview of a sequence or a tree, use the **Fit Width** function instead of the **Zoom Out** function.

If you press **Shift** while clicking in a **View**, the zoom funtion is reversed. Hence, clicking on a sequence in this way while the **Zoom Out** mode toolbar item is selected, zooms in instead of zooming out.

3.3.3 Fit Width

The **Fith Width** ( ) function adjusts the content of the **View** so that both ends of the sequence, alignment, or tree is visible in the **View** in question. (This function does not change the mode of the mouse pointer.)

3.3.4 Zoom to 100%

The **Zoom to 100%** ( ) function zooms the content of the **View** so that it is displayed with the highest degree of detail. (This function does not change the mode of the mouse pointer.)

3.3.5 Move

The Move mode allows you to drag the content of a **View**. E.g. if you are studying a sequence, you can click anywhere in the sequence and hold the mouse button. By moving the mouse you move the sequence in the **View**.

3.3.6 Selection

The Selection mode ( ) is used for selecting in a **View** (selecting a part of a sequence, selecting nodes in a tree etc.). It is also used for moving e.g. branches in a tree or sequences in an alignment.

When you make a selection on a sequence or in an alignment, the location is shown in the bottom right corner of your workbench. E.g. '23^24' means that the selection is between two residues. '23' means that the residue at position 23 is selected, and finally '23..25' means that 23, 24 and 25 are selected. By holding ctrl / ⌘ you can make multiple selections.

3.4 Toolbox and Status Bar

The **Toolbox** is placed in the left side of the user interface of *CLC Gene Workbench 2.0* below the **Navigation Area**.

The **Toolbox** shows a **Processes** tab and a **Toolbox** tab.

3.4.1 Processes

By clicking the **Processes** tab, the **Toolbox** displays previous and running processes, e.g. an NCBI search or a calculation of an alignment. The running processes can be stopped, paused, and resumed.

Active buttons are blue.

If a process is terminated, the stop, pause, and play buttons of the process in question are made gray.

The terminated processes can be removed by:

View | **Remove Terminated Processes (** **)**

Running and paused processes are not deleted.

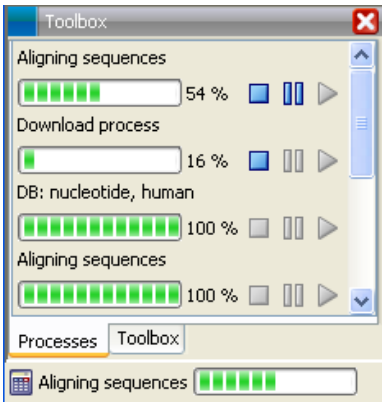


Figure 3.14: *Two running, and a number of terminated processes in the Toolbox.*

If you close the program while there are running processes, a dialog will ask if you are sure that you want to close the program. Closing the program will stop the process, and it cannot be restarted when you open the program again.

3.4.2 Toolbox

The content of the **Toolbox** tab in the **Toolbox** corresponds to **Toolbox** in the **Menu Bar**.

The **Toolbox** can be hidden, so that the **Navigation Area** is enlarged and thereby displays more elements:

View | **Show/Hide Toolbox**

The tools in the toolbox can be accessed by double-clicking or by dragging elements from the **Navigation Area** to an item in the **Toolbox**.

3.4.3 Status Bar

As can be seen from figure 3.1, the **Status Bar** is located at the bottom of the window. In the left side of the bar is an indication of whether the computer is making calculations or whether it is idle. The right side of the **Status Bar** indicates the range of the selection of a sequence. (See chapter 3.3.6 for more about the Selection mode button.)

3.5 Workspace

If you are working on a project and have arranged the views for this project, you can save this arrangement using **Workspaces**. A Workspace remembers the way you have arranged the views, and you can switch between different workspaces.

The **Navigation Area** always contains the same data across **Workspaces**. It is, however, possible to open different folders in the different **Workspaces**. Consequently, the program allows you to display different clusters of the data in separate **Workspaces**.

All **Workspaces** are automatically saved when closing down *CLC Gene Workbench 2.0*. The next time you run the program, the **Workspaces** are reopened exactly as you left them.

Notice! It is not possible to run more than one version of *CLC Gene Workbench 2.0* at a time. Use two or more **Workspaces** instead.

3.5.1 Create Workspace

When working with large amounts of data, it might be a good idea to split the work into two or more **Workspaces**. As default the *CLC Gene Workbench* opens one **Workspace**, (the largest window in the right side of the workbench, see 3.1). Additional **Workspaces** are created in the following way:

Workspace in the Menu Bar) | **Create Workspace** | **enter name of Workspace** | **OK**

When the new **Workspace** is created, the heading of the program frame displays the name of the new **Workspace**. Initially, the **Project Tree** in the **Navigation Area** is collapsed and the **View Area** is empty and ready to work with. (See figure 3.15).

3.5.2 Select Workspace

When there is more than one **Workspace** in the workbench, there are two ways to switch between them:

Workspace ( ) in the Toolbar | **Select the Workspace to activate**

or **Workspace in the Menu Bar** | **Select Workspace ( )** | **choose which Workspace to activate** | **OK**

The name of the selected **Workspace** is shown after "*CLC Gene Workbench 2.0*" at the top left corner of the main window, in this case: (default).

3.5.3 Delete Workspace

Deleting a **Workspace** can be done in the following way:

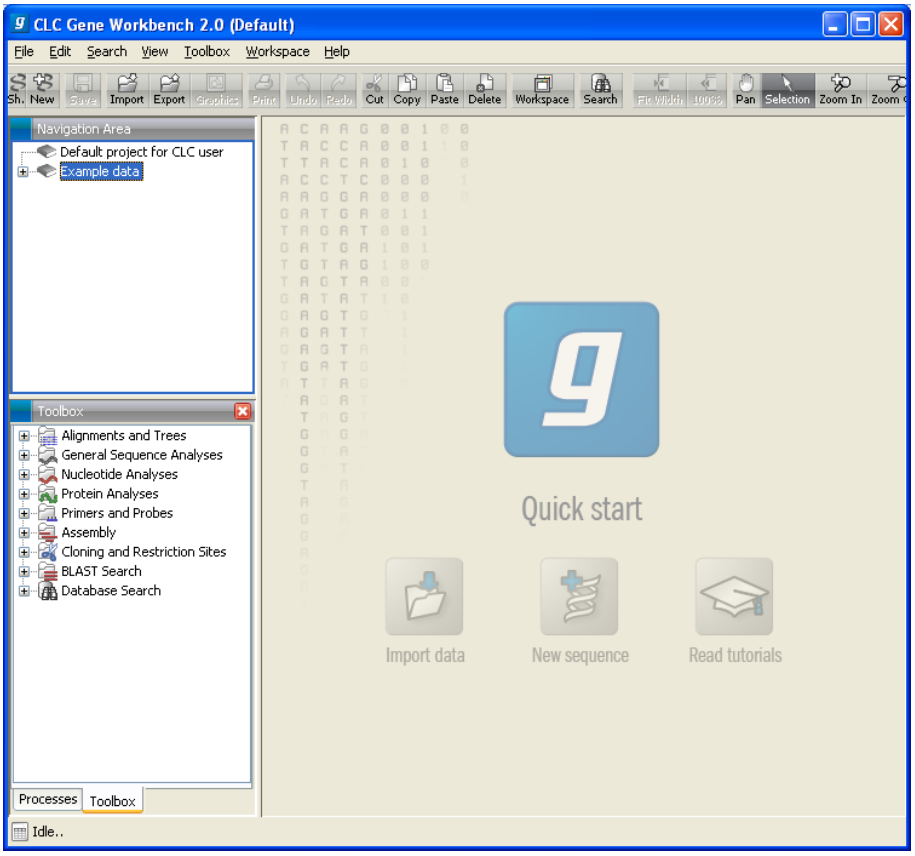


Figure 3.15: *An empty Workspace.*

Workspace in the Menu Bar | Delete Workspace | choose which Workspace to delete | OK

Notice! Be careful to select the right **Workspace** when deleting. The delete action cannot be undone. (However, no data is lost, because a workspace is only a representation of data.)

It is not possible to delete the default workspace.

3.6 List of shortcuts

The keyboard shortcuts in *CLC Gene Workbench 2.0* are listed below.

Action	Windows/Linux	Mac OS X
Adjust selection	Shift + arrow keys	Shift + arrow keys
Change between tabs	Ctrl + tab	⌘ + tab
Close	Ctrl + W	⌘ + W
Close all views	Ctrl + Shift + W	⌘ + Shift + W
Copy	Ctrl + C	⌘ + C
Cut	Ctrl + X	⌘ + X
Delete	Delete	Delete
Exit	Alt + F4	⌘ + Q
Export	Ctrl + E	⌘ + E
Export graphics	Ctrl + G	⌘ + G
Find Inconsistency	Space	Space
Find Previous Inconsistency	,	,
Help	F1	F1
Import	Ctrl + I	⌘ + I
Maximize/restore size of View	Ctrl + M	⌘ + M
Move gaps in alignment	Ctrl + arrow keys	⌘ + arrow keys
Navigate sequence views	left/right arrow keys	left/right arrow keys
New Folder	Ctrl + Shift + N	⌘ + Shift + N
New Project	Ctrl + R	⌘ + R
New Sequence	Ctrl + N	⌘ + N
View	Ctrl + O	⌘ + O
Paste	Ctrl + V	⌘ + V
Print	Ctrl + P	⌘ + P
Redo	Ctrl + Y	⌘ + Y
Rename	F2	F2
Save	Ctrl + S	⌘ + S
Search in an open sequence	Ctrl + F	⌘ + F
Search NCBI	Ctrl + B	⌘ + B
Search UniProt	Ctrl + Shift + U	⌘ + Shift + U
Select All	Ctrl + A	⌘ + A
Selection Mode	Ctrl + 2	⌘ + 2
User Preferences	Ctrl + K	⌘ + ;
Split Horizontally	Ctrl + T	⌘ + T
Split Vertically	Ctrl + J	⌘ + J
Show/hide Preferences	Ctrl + U	⌘ + U
Undo	Ctrl + Z	⌘ + Z
Zoom In Mode	Ctrl + + (plus)	⌘ + + (plus)
Zoom In (without clicking)	+ (plus)	+ (plus)
Zoom Out Mode	Ctrl + - (minus)	⌘ + - (minus)
Zoom Out (without clicking)	- (minus)	- (minus)

Combinations of keys and mouse movements are listed below.

Action	Windows/Linux	Mac OS X	Mouse movement
Maximize View			Double-click the tab of the View
Restore View			Double-click the View title
Reverse zoom function	Shift	Shift	Click in view
Select multiple elements	Ctrl	⌘	Click elements
Select multiple elements	Shift	Shift	Click elements

Chapter 4

User preferences

Contents

4.1 General preferences 77
4.2 Default View preferences 77
4.3 Advanced preferences 78
4.4 Export/import of preferences 78
4.5 View preference style sheet 78
4.5.1 Floating Side Panel 79

The **Preferences** dialog offers opportunities for changing the default settings for different features of the program. For example, if you adjust **Number of hits** under **General Preferences** to 40 (instead of 50), you see the first 40 hits each time you conduct a search (e.g. NCBI search).

The **Preferences** dialog is opened in one of the following ways and can be seen in figure 4.1:

Edit | **Preferences (** **)**

or **Ctrl + K (⌘ + ; on Mac)**

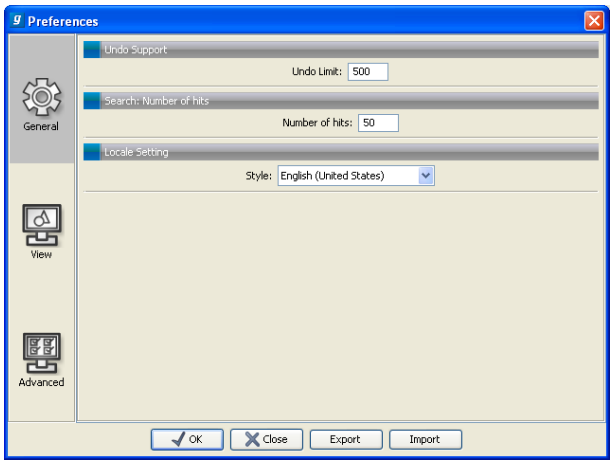


Figure 4.1: *Preferences include General preferences, View preferences, Colors preferences, and Advanced settings.*

4.1 General preferences

The **General** preferences include:

- **Undo Limit.** As default the undo limit is set to 500. By writing a higher number in this field, more actions can be undone. Undo applies to all changes made on sequences, alignments or trees. See section 3.2.4 for more on this topic.
- **Number of hits.** The number of hits shown in *CLC Gene Workbench 2.0*, when e.g. searching NCBI. (The sequences shown in the program are not downloaded, until dragged/saved into the Navigation Area.
- **Locale Setting.** i.e. in which country you are located. This determines the punctuation to be used.

4.2 Default View preferences

There are five groups of default **View** settings:

1. **Toolbar**
2. **Side Panel Location**
3. **New View**
4. **View Format**
5. **Default view settings sheet**.

In general, these are default settings for the user interface.

The fToolbar preferences let you choose the size of the toolbar icons, and you can choose whether to display names below the icons.

The **Side Panel Location** setting lets you choose between **Dock in views** and **Float in window**. When docked in view, view preferences will be located in the right side of the view of e.g. an alignment. When floating in window, the side panel can be placed everywhere in your screen, also outside the workspace, e.g. on a different screen. See section 4.5 for more about floating side panels.

The **New view** setting allows you to choose whether the **View preferences** are to be shown automatically when opening a new view. If this option is not chosen, you can press (Ctrl + U (⌘ + U on Mac)) to see the preferences panels of an open view.

The **View Format** allows you to change the way the elements appear in the **Navigation Area**. The following text can be used to describe the element:

- Name (this is the default information to be shown).
- Accession (sequences downloaded from databases like GenBank have an accession number).
- Species.

- Species (accession).
- Common Species.
- Common Species (accession).

The **User Defined View Settings** gives you an overview of different style sheets for your **View preferences**. See section 4.5 for more about how to create and save style sheets.

The first time you use the program, only the **CLC Standard Settings** is available. However, the tab allowing you to choose the style sheet for a viewer (e.g. a sequence viewer) only appears after you have launched the viewer for the first time.

4.3 Advanced preferences

The **Advanced** settings include the possibility to set up a proxy server. This is described in section 1.7 .

4.4 Export/import of preferences

The user preferences of the *CLC Gene Workbench 2.0* can be exported to other users of the program, allowing other users to display data with the same preferences as yours. You can also use the export/import preferences function to backup your preferences.

To export preferences, open the **Preferences** dialog (Ctrl + K (⌘ + ; on Mac)) and do the following:

> **Export** | **Select the relevant preferences** | **Export** | **Choose location for the exported file** | **Enter name of file** | **Save**

Notice! The format of exported preferences is .cpf. This notation must be submitted to the name of the exported file in order for the exported file to work.

Notice! Before exporting, you are asked about which of the different settings you want to include in the exported file. "Default View Settings Sheet", which is one of the preferences which can be selected for export, does not include the Style sheets themselves, but only information about which of the Style sheets is default style sheets.

The process of importing preferences is similar to exporting:

> **Press Ctrl + K (⌘ + ; on Mac) to open Preferences** | **Import** | **Browse to and select the .cpf file** | **Import and apply preferences**

4.5 View preference style sheet

Depending on which view you have opened in the Workbench, you have different options of adjusting the **View preferences**.

Figure 4.2 shows the preference groups which are available for a sequence.

By clicking the black triangles, the different preference groups can be opened. An example is shown in figure 4.3.

Figure 4.2: *View preferences for a view of a sequence include several preference groups. In this case the groups are: Sequence layout, Annotation types, Annotation layout, etc. Several of these preference groups are present in more views. E.g. Sequence layout is also present when an alignment is viewed.*

The content of the different preference groups, are described in connection to those chapters where the functionality is explained. E.g. **Sequence Layout** View preferences are described in chapter 11.1.1 which is about editing options of a sequence view.

When you have adjusted a view of e.g. a sequence, your settings can be saved in a so called style sheet. When you open other sequences, which you want to display in a similar way, the saved style sheet can be applied. These options are available in the top of the View preferences. (See figure 4.4).

To manage style sheets click ( ) seen in figure 4.4. This opens a menu, where the following options are available:

- Save Settings
- Delete Settings
- Apply Saved Settings

Style sheets for the View preference differ between views. Hence, you can have e.g. three style sheets for sequences, two for alignments, and four for graphs. To adjust which of the style sheets is default for e.g. an alignment, go to the general **Preferences** (Ctrl + K (⌘ + ; on Mac).

CLC Standard Settings represents the way the program was set up, when you first launched the program.

The remaining icons of figure 4.4 are used to; **Expand all preferences**, **Collapse all preferences**, and **Dock/Undock Preferences**. **Dock/Undock Preferences** is used when making the View preferences "floating". See next section

4.5.1 Floating Side Panel

The Side Panel of the views can be placed in the right side of a view, or they can be floating. (See figure 4.5).

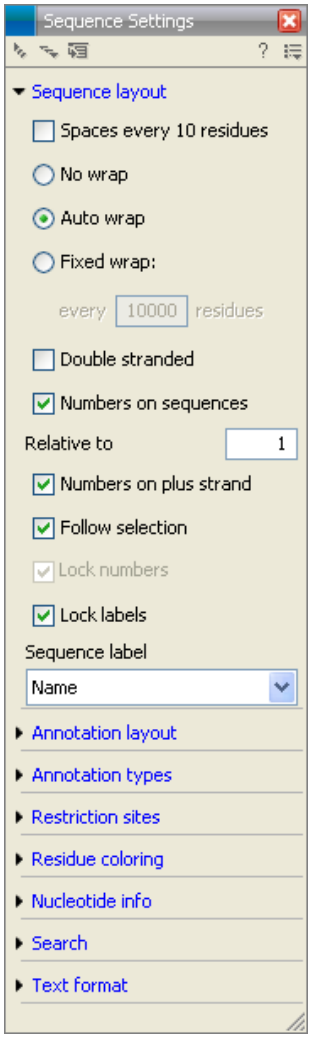


Figure 4.3: *The many preferences for each view are stored in preference groups which can be opened and closed.*

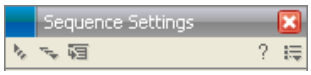


Figure 4.4: *The top of the View preferences contain Expand all preferences , Collapse all preferences, Dock/Undock preferences, Help, and Save/Restore preferences.*

By clicking the Dock icon ( ) the floating Side Panel reappear in the right side of the view. The size of the floating Side Panel can be adjusted by dragging the hatched area in the bottom right.

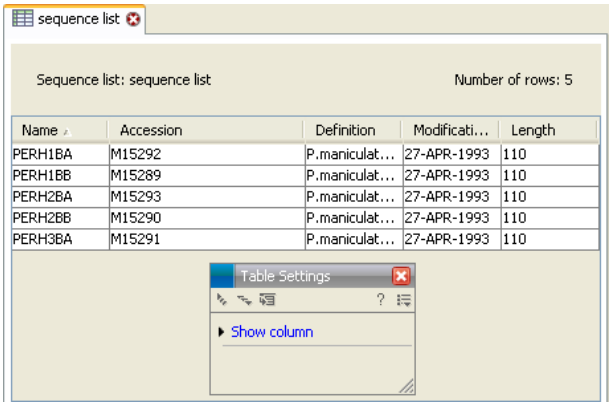


Figure 4.5: *The floating Side Panel can be moved out of the way, e.g. to allow for a wider view of a table.*

Chapter 5

Printing

Contents

5.1 Selecting which part of the view to print 82
5.2 Page setup 83
5.3 Print preview 83

CLC Gene Workbench 2.0 offers different choices of printing the result of your work.

This chapter deals with printing directly from the workbench. Another option for using the graphical output of your work, is to export graphics (see chapter 6.3) in a graphic format, and then import it into a document or into a presentation.

All the kinds of data that you can view in the **View Area** can be printed. For some of the views, the layout will be slightly changed in order to be printer-friendly.

It is not possible to print elements directly from the **Navigation Area**. They must first be opened in a view in order to be printed:

select relevant view | Print ( ) in the toolbar

If you are printing e.g. alignments, sequences and graphs, you will be faced with three different dialogs, allowing you to adjust the way your view is printed.

- A dialog to let you select which part of the view you want to print.
- A dialog to adjust page setup.
- A **Print preview** window.

These three kinds of dialogs are described in the two following sections.

5.1 Selecting which part of the view to print

Views that are printed exactly like they look on the screen, have an option for selecting which part of the view to print (see figure 5.1).

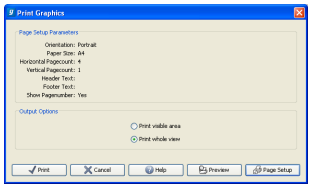


Figure 5.1: *When printing graphics you get the options of printing the visible area or printing the whole view.*

Printing the whole view is useful if you have zoomed in on an area of the view, and you want to print the whole view (also the part of e.g. a sequence, which is not visible). On the other hand, if you want to print some details of an area of the view, you can use the zoom and navigate functions first, and then print the visible area. This will result in a print of only some part of the sequence.

5.2 Page setup

No matter whether you have chosen to print the visible area or the whole view, you can adjust page setup of the print. An example of this can be seen in figure 5.2

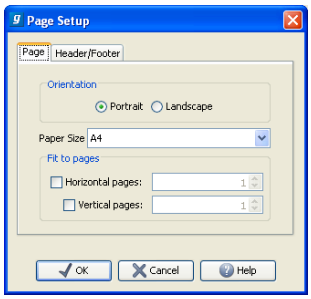


Figure 5.2: *In this dialog the default settings Portrait and A4 apply to print of an alignment. By checking Fit to pages it is possible to adjust Horizontal pages to 2. This is done allow a long sequence to stretch the width of two A4 pages. This is illustrated in the Page Layout field.*

Click the **Header/Footer** tab to edit the header and footer text. By clicking in the text field for either **Custom header text** or **Custom footer text** you can access the auto formats for header/footer text in **Insert a caret position**. Click either **Date**, **View name**, or **User name** to include the auto format in the header/footer text.

Click **OK** to see the print preview with the settings you have made.

5.3 Print preview

The preview is shown in figure 5.3).

The **Print preview** window lets you see the layout of the pages that are printed. Use the arrows in the toolbar to navigate between the pages. Click Print ( ) to show the print dialog, which lets you choose e.g. which pages to print.

Notice that if you wish to change e.g. the colors of the residues in the alignment, this must be changed in the **View preferences** of the specific dot plot.

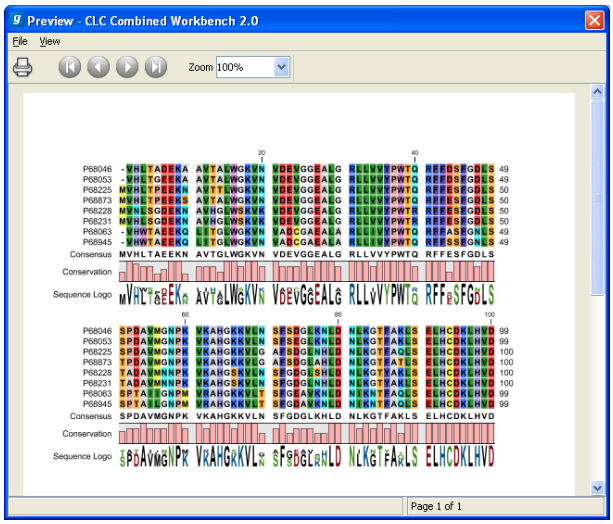


Figure 5.3: *Print preview.*

Chapter 6

Import/export of data and graphics

Contents

6.1 Bioinformatic data formats 85
6.1.1 Import of bioinformatic data 86
6.1.2 Export of bioinformatic data 88
6.2 External files 90
6.2.1 Import external files 90
6.2.2 Export external files 90
6.2.3 Technical details 91
6.3 Export graphics to files 91
6.3.1 Exporting protein reports 93
6.4 Copy/paste view output 93

CLC Gene Workbench 2.0 handels a large number of different data formats. All data stored in the Workbench is available in the **Navigation Area** of the program. The data of the **Navigation Area** can be divided into two groups. The data is either one of the different bioinformatic data formats, or it can be an 'external file'. Bioinformatic data formats are those formats which the program can work with, e.g. sequences, alignments and phylogenetic trees. External files are files or links which are stored in *CLC Gene Workbench 2.0*, but are opened by other applications, e.g. pdf-files, Microsoft Word files, Open Office spreadsheet files, or it could be links to programs and webpages etc.

Furthermore, this chapter deals with the export of graphics.

6.1 Bioinformatic data formats

The different bioinformatic data formats are imported in the same way, therefore, the following description of data import is an example which illustrates the general steps to be followed, regardless of which format you are handling.

6.1.1 Import of bioinformatic data

Here follows a short list of the formats which *CLC Gene Workbench 2.0* handles, and a description of which type of data the different formats support.

File type	Suffix	File format used for
Phylip Alignment	.phy	alignments
GCG Alignment	.msf	alignments
Clustal Alignment	.aln	alignments
Newick	.nwk	trees
FASTA	.fsa/.fasta	sequences
GenBank	.gbk/.gb/.gp	sequences
GCG sequence	.gcg	sequences (only import)
PIR (NBRF)	.pir	sequences (only import)
Staden	.sdn	sequences (only import)
VectorNTI		sequences (only import)
DNAstrider	.str/.strider	sequences
Swiss-Prot	.swp	protein sequences
Lasergene sequence	.pro	protein sequence (only import)
Lasergene sequence	.seq	nucleotide sequence (only import)
Embl	.embl	nucleotide sequences
Nexus	.nxs/.nexus	sequences, trees, alignments, and sequence lists
CLC	.clc	sequences, trees, alignments, reports, etc.
Text	.txt	all data in a textual format
ABI		Trace files (only import)
AB1		Trace files (only import)
SCF2		Trace files (only import)
SCF3		Trace files (only import)
Phred		Trace files (only import)
mmCIF	.cif	structure (only import)
PDB	.pdb	structure (only import)
Preferences	.cpf	CLC workbench preferences

Notice that *CLC Gene Workbench* can import 'external' files, too. This means that *CLC Gene Workbench* can import all files and display them in the **Navigation Area**, while the above mentioned formats are the types which can be read by *CLC Gene Workbench*.

The *CLC Gene Workbench 2.0* offers a lot of possibilities to handle bioinformatic data. Read the next sections to get information on how to import different file formats or to import data from a Vector NTI database.

Import of common bioinformatic data

Before importing a file, you must decide where you want to import it, i.e. which project or folder. The imported file ends up in the project or folder you selected in the **Navigation Area**.

> **select project or folder** | **click Import ( ) in the Toolbar** | **browse to the relevant file** | **Select**

The imported file is placed at the location which was selected when the import was initiated. E.g. if you right-click on a file in the **Navigation Area** and choose import, the imported file is placed

immediately below the selected file. If you right-click a folder, the imported file is placed as the last file in that folder. If you right-click a project, the imported file is placed as the last file in that project (and after existing folders).

It is also possible to drag a file from e.g. the desktop into the **Navigation Area** of *CLC Gene Workbench*. If *CLC Gene Workbench* recognizes the file format, the file is automatically parsed (changed) into CLC format and stored in the **Navigation Area**. If the format is not recognized, the following dialog is displayed (see figure 6.1):

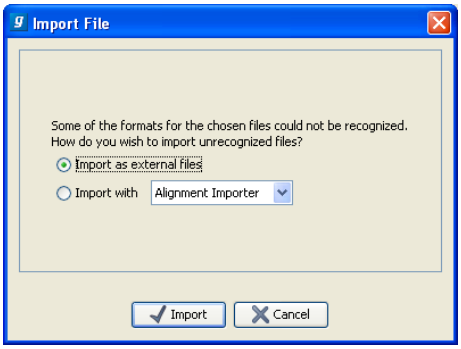

Import File
Some of the formats for the chosen files could not be recognized.
How do you wish to import unrecognized files?
Import as external files
Import with
Alignment Importer
Import
Cancel

Figure 6.1: *If the dragged file is not recognized by* **CLC Gene Workbench** *the dialog allows you to "force" the import in a certain format.*

Notice! When browsing for files to import, the dialog only displays files of the format chosen in the **File of type** drop down menu at the bottom of the import dialog. If the format .clc is chosen, only .clc-files are shown in the **Import** dialog. Choose **All Files** to ensure the file you are looking for is displayed.

When you import a file containing several sequences, you will be asked whether you want to save the sequences as individual elements or as a sequence list (see section 11.5 for more about sequence lists).

Import of data in clc-format from older versions

If you want to import data in clc-format generated in an older version of either of the workbenches, it has to bee converted first. If you try to import it without conversion, you will see a warning dialog.

Import of Vector NTI data

CLC Gene Workbench 2.0 can import DNA, RNA, and protein sequences from a Vector NTI Database. The import can be done for Vector NTI Advance^{TM}10 for Windows machines and Vector NTI Suite 7.1 for Mac OS X for Panther and former versions. A new Project will be placed in the **Navigation Area** and you can find all sequences in different folders ready to work with. In order to import all DNA/RNA and protein sequences:

> **select File in the Menu Bar** | **Import VectorNTI Data..** | **select a database directory** | **Import** | **confirm the information**

Notice! The default installation of the VectorNTI program for the database home is

- C:/VNTI Database/
  for Windows machines and

- /Library/Application Support/VNTI Database/
  for Mac OS X for Panther.

Therefore the *CLC Gene Workbench 2.0* will check if there is a default installation and will ask whether you want to use the default database directory or another directory.

Notice! Make sure that the Vector NTI database directory (default or backup) contains folders like ProData and MolData. These folders are necessary when we import the data into *CLC Gene Workbench 2.0*.

In order to import all DNA/RNA and protein sequences if a default database directory is installed:

select File in the Menu Bar | Import VectorNTI Data | select Yes if you want to import the default database | confirm the information

or **select File in the Menu Bar | Import VectorNTI Data | select No to choose a database | select a database directory | Import | confirm the information**

After the import there is a new Project called **Vector NTI Data** in the **Navigation Area**. In **Vector NTI Data** you can see two folders: **DNA/RNA** containing the DNA and RNA sequences, and **Protein** containing all protein sequences. (See figure 6.2).

The project, folders and all sequences are automatically saved.

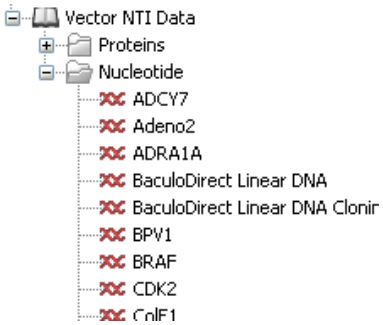


Figure 6.2: *Project Vector NTI Data containing all imported sequences of the Vector NTI Database.*

6.1.2 Export of bioinformatic data

CLC Gene Workbench 2.0 can export bioinformatic data in most of the formats that can be imported. There are a few exceptions. See section 6.1.1.

To export a file:

select the element to export | Export ( ) | choose where to export to | select 'File of type' | enter name of file | Save

Notice! The **Export** dialog decides which types of files you are allowed to export into, depending on what type of data you want to export. E.g. protein sequences can be exported into GenBank, Fasta, Swiss-Prot and CLC-formats.

Export of projects, folders and multiple files

The .clc file type can be used to export all kinds of files and is therefore especially useful in these situations:

- Export of one or more file folders including all underlying files and folders.
- Export of one or more project folders including all underlying files and folders.
- If you want to export two or more files into one .clc-file, you have to copy them into a folder or project, which can be exported as described below:

Export of projects and folders is similar to export of single files. Exporting multiple files (of different formats) is done in .clc-format. This is how you export a project:

> **select the project to export** | **Export ( )** | **choose where to export to** | **enter name of project** | **Save**

You can export multiple files of the same type into formats other than CLC (.clc). E.g. two DNA sequences can be exported in GenBank format:

> **select the elements to export by <Ctrl>-click or <Shift>-click** | **Export ( )** | **choose where to export to** | **choose GenBank (.gbk) format** | **enter name of project** | **Save**

Export of dependent objects

When exporting e.g. an alignment, *CLC Gene Workbench 2.0* can export all dependent objects. I.e. the sequences which the alignment is calculated from. This way, when sending your alignment (with the dependent objects), your colleagues can reproduce your findings with adjusted parameters, if desired.
To export with dependent files:

> **select the element in Navigation Area** | **File in Menu Bar** | **Export with dependent objects** | **enter name of project** | **choose where to export to** | **Save**

The result is a folder containing the exported file with dependent objects, stored automatically in a folder on the desired location of your desk.

Export history

To export an element's history:

> **select the element in Navigation Area Export( )** | **select History PDF(.pdf)** | **choose where to export to** | **Save**

The entire history of the element is then exported in pdf format.

The CLC format

CLC Gene Workbench keeps all bioinformatic data in the CLC format. Compared to other formats, the CLC format contains more information about the object, like its history and comments. The CLC format is also able to hold several objects of different types (e.g. an alignment, a graph and a phylogenetic tree). This means that if you are exporting your data to another CLC Workbench, you can use the CLC format to export several objects in one file, and all the objects' information is preserved.

Notice! CLC files can be exported from and imported into all the different CLC Workbenches.

Back up

The CLC format is practical for making manual back up of your files. All files are stored in Projects and these can easily be exported out of *CLC Gene Workbench*, :

> **select the project to export** | **Export ( )** | **choose where to export to** | **enter name of project** | **Save**

Other than that, the files of the **Navigation Area** are stored in a persistence folder on your computer. Hence, your regular back up system should be set up to include this folder.

On Mac the folder can be found: Library/Application Support/CLC bio/Workbench/<version number>/persistence

On Windows: Documents and Settings/<username>/CLC bio/Workbench/<version number>/persistence

On Linux: home/<username>/.clcbio/workbench/<version number>/persistence

6.2 External files

In order to help you organize your projects, *CLC Gene Workbench 2.0* lets you import all kinds of files. E.g. if you have Word, Excel or pdf-files related to your project, you can import them into a project in *CLC Gene Workbench 2.0*. Importing an external file creates a copy of the file which is saved in a project in *CLC Gene Workbench 2.0*. The file can now be opened by double-clicking the file name in the **Navigation Area**. The file is opened using the default application for this file type (e.g. Microsoft Word for .doc-files and Adobe Reader for .pdf).

CLC Gene Workbench can also show web links (URLs) in the **Navigation Area**. This can be done by using the **Import** function of the program or by dragging the file e.g. from the desktop to the **Navigation Area**.

6.2.1 Import external files

To import an external file:

> **click a project or folder to import into** | **Import ( ) in the toolbar** | **Choose All files in Files of type** | **browse to the relevant file** | **Select**

or **drag the file from the file system into a project in the Navigation Area (only possible under Windows)**

Notice! When you import an external file, a copy of the original file is created. This means that you should always make sure that you open the file from within *CLC Gene Workbench 2.0*.

6.2.2 Export external files

If you export an entire project or folder from *CLC Gene Workbench 2.0*, the exported CLC file will include all external files stored in it. This means that you can export the project as a CLC file, and send it to a colleague who can import it and access all the files in the project.

You can also export individual files in their original format. To export a file from *CLC Gene Workbench 2.0*:

click a file in the Navigation Area | Export ( ) in the toolbar | browse to the desired folder | Save

If the file already exists, you are asked if you want to replace it.

6.2.3 Technical details

This section explains the more technical aspects of how *CLC Gene Workbench 2.0* stores the external files. When you import the file, a copy of the file is created in a database. When you open the file from the **Navigation Area**, it's checked out to a repository (a folder called "CLCWorkbenchRepository" located in your operating system's user folder) where it stays until you close the application that has the file open. When you exit *CLC Gene Workbench 2.0*, it checks all the files in the repository into the database, unless they are still open in another application. If the latter is the case, the file stays in the repository even after the file is closed, and it will not be checked in until the next time *CLC Gene Workbench 2.0* is closed.

If you have made changes to a file after the *CLC Gene Workbench 2.0* was closed, a dialog is shown asking which version to use. The date and time of the latest change of the file is displayed in the dialog helping you to decide which one to keep (see figure 6.3).

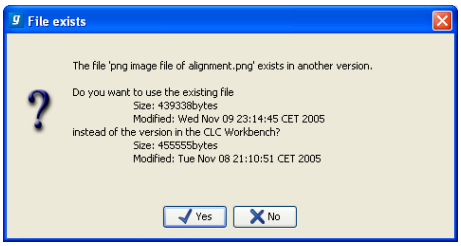

File exists
The file 'png image file of alignment.png' exists in another version.
Do you want to use the existing file
Size: 439338bytes
Modified: Wed Nov 09 23:14:45 CET 2005
instead of the version in the CLC Workbench?
Size: 455555bytes
Modified: Tue Nov 08 21:10:51 CET 2005
Yes
No

Figure 6.3: *A dialog asking which version of the file you want to keep.*

6.3 Export graphics to files

CLC Gene Workbench 2.0 supports export of graphics into a number of formats. This way, the visible output of your work can easily be saved and used in presentations reports etc. The **Export Graphics** function ( ) is found in the **Toolbar**.

CLC Gene Workbench 2.0 exports graphics exactly the way it is shown in the **View Area**. Thus, all settings made in the **Side Panel** will be reflected in the exported file.

To show you how to export graphics, we choose to export the phylogenetic tree of the example data set in .png-format. See 6.4.

When the relevant file is opened and shown in the **View Area** do the following:

select tab of View | Graphics ( ) on Toolbar | select location on disc | name file and select type | Save

After clicking **Save**, you are prompted for whether to **Export visible area** or **Export whole view**. The first parameter exports 'what you see' and the latter parameter also exports the part of the view that is not visible. Hence, choosing **Export whole view** will generate a larger file.

Furthermore, when saving in .png, .jpg, and .tif-formats you are prompted for which quality to save the graphics in.

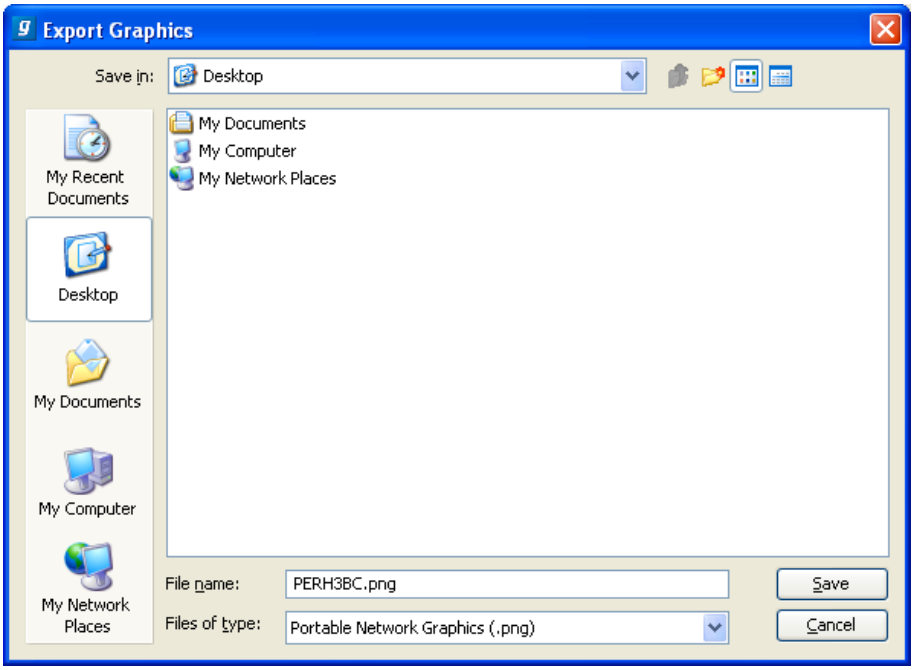


Figure 6.4: *Exporting a phylogenetic tree from* **CLC Gene Workbench 2.0**.

To see the exported file browse to the file on your computer and open it. In our case the .png-file is opened in a browser, the result can be seen in figure 6.5.

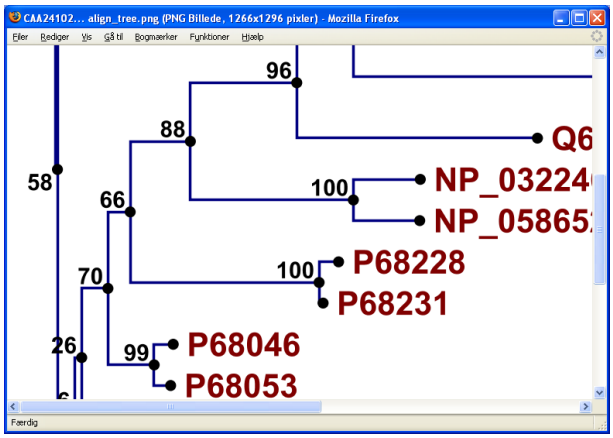


Figure 6.5: *The exported .png-file opened in a browser. (Due to high resolution of the exported graphics, it is not possible to see the entire file in the browser window.)*

The following file types are available for exporting graphics in *CLC Gene Workbench 2.0*:

Bitmap images
In a bitmap image, each dot in the image has a specified color. This implies, that if you zoom in on the image there will not be enough dots, and if you zoom out there will be too many. In these cases the image viewer has to interpolate the colors to fit what is actually looked at. This format is a good choice for storing images without large shapes (e.g. dot plots).

Vector graphics
Vector graphics is a collection of shapes. Thus what is stored is e.g. information about where a line starts and ends, and the color of the line and its width. This enables a given viewer to decide how to draw the line, no matter what the zoomfactor is, thereby always giving a correct

Format	Suffix	Type
Portable Network Graphics	.png	bitmap
JPEG	.jpg	bitmap
Tagged Image File	.tif	bitmap
PostScript	.ps	vector graphics
Encapsulated PostScript	.eps	vector graphics
Portable Document Format	.pdf	vector graphics
Scalable Vector Graphics	.svg	vector graphics

image. This format is good for e.g. graphs and reports, but less usable for e.g. dotplots.

Graphics files can also be imported into the **Navigation Area**. However, no kinds of graphics files can be displayed in *CLC Gene Workbench 2.0*. See section 6.2.1 for more about importing external files into *CLC Gene Workbench 2.0*.

6.3.1 Exporting protein reports

Protein reports cannot be exported in the same way as other data. Instead, they can be exported from the **Navigation Area**:

Click the report in the Navigation Area | **Export ( ) in the Toolbar** | **select pdf**

When the report is exported, the file can be opened with Adobe Reader. Opening and printing in Adobe Reader is also the only way to print the report.

6.4 Copy/paste view output

The content of tables, e.g. in reports, folder lists, and sequence lists can be copy/pasted into different programs, where it can be edited. *CLC Gene Workbench 2.0* pastes the data in tabulator separated format which is useful if you use programs like Microsoft Word and Excel. There is a huge number of programs in which the copy/paste can be applied. For simplicity, we include one example of the copy/paste function from a **Folder Content** view to Microsoft Excel.

First step is to select the desired elements in the view:

click a line in the Folder Content view | **hold Shift-button** | **Push arrow down (or up)**

See figure 6.6.

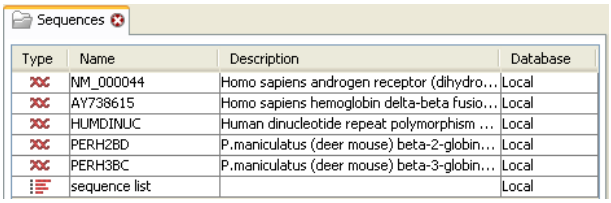


Figure 6.6: *Selected elements in a Folder Content view.*

When the elements are selected, do the following to copy the selected elements:

right-click one of the selected elements | **Edit** | **Copy ( )**

Then:

right-click in the cell A1 | Paste ( )

The outcome might appear unorganized, but with a few operations the structure of the view in *CLC Gene Workbench 2.0* can be produced. (Except the icons which are replaced by file references in Excel.)

Chapter 7

History

Contents

7.1 Element history 95
7.1.1 Sharing data with history 96

CLC Gene Workbench 2.0 keeps a log of all operations you make in the program. If e.g. you rename a sequence, align sequences, create a phylogenetic tree or translate a sequence, you can always go back and check what you have done. In this way, you are able to document and reproduce previous operations.

This can be useful in several situations: It can be used for documentation purposes, where you can specify exactly how your data has been created and modified. It can also be useful if you return to a project after some time and want to refresh your memory on how the data was created. Also, if you have performed an analysis and you want to reproduce the analysis on another element, you can check the history of the analysis which will give you all parameters you set.

This chapter will describe how to use the **History** functionality of *CLC Gene Workbench 2.0*.

7.1 Element history

You can view the history of all elements in the **Navigation Area** except files that are opened in other programs (e.g. Word and pdf-files). The history starts when the element appears for the first time in *CLC Gene Workbench 2.0*. To view the history of an element:

Right-click the element in the Navigation Area | **Show** | **History ( )**

or **Select the element in the Navigation Area** | **Show ( ) in the Toolbar** | **History ( )**

This opens a view that looks like the one in figure 7.1.

When opening an element's history is opened, the newest change is submitted in the top of the view. The following information is available:

- **Title**. The action that the user performed.

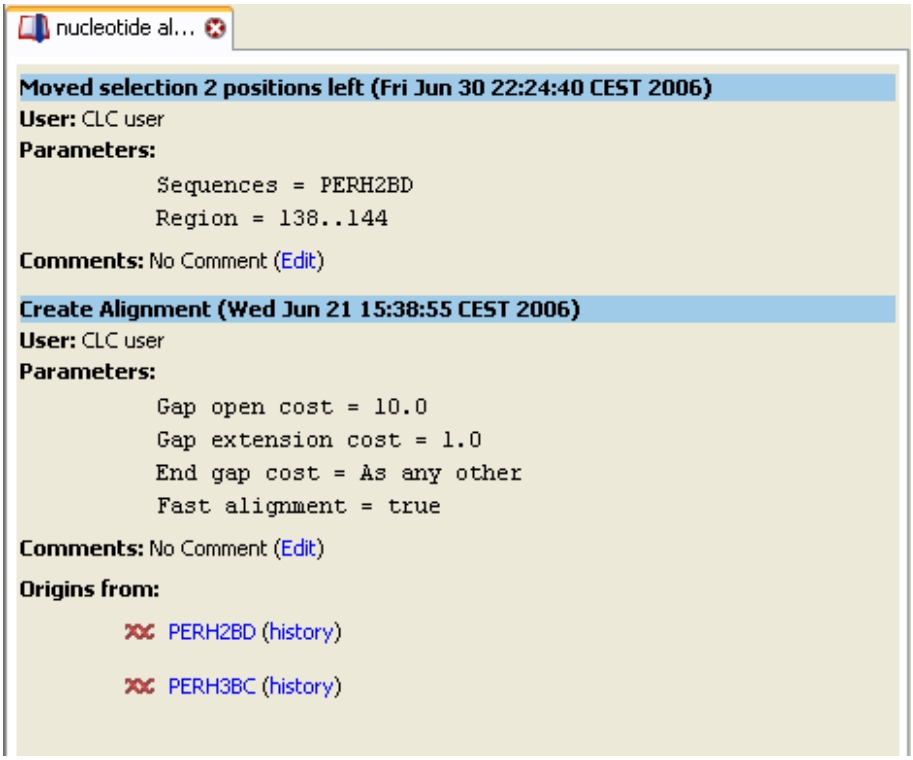


Figure 7.1: *An element's history.*

- **Date and time**. Date and time for the operation. The date and time are displayed according to your locale settings (see section 4.1).
- **User**. The user who performed the operation. If you import some data created by another person in a CLC Workbench, that persons name will be shown.
- **Parameters**. Details about the action performed. This could be the parameters that was chosen for an analysis.
- **Origins from**. This information is usually shown at the bottom of an element's history. Here, you can see which elements the current element origins from. If you have e.g. created an alignment of three sequences, the three sequences are shown here. Clicking the element selects it in the **Navigation Area**, and clicking the 'history' link opens the element's own history.

7.1.1 Sharing data with history

The history of an element is attached to that element, which means that exporting an element in CLC format (*.clc) will export the history too. In this way, you can share projects and files with others while preserving the history. If an element's history includes source elements (i.e. if there are elements listed in 'Origins from'), they must also be exported in order to see the full history. Otherwise, the history will have entries named "Element deleted". An easy way to export an element with all its source elements is to use the **Export Dependent Objects** function described in section 6.1.2.

The of a history view can be printed. To do so, click the **Print** icon ( ).

Chapter 8

Handling of results

Contents

8.1 How to handle results of analyses 97
8.1.1 When the analysis does not create new elements 97
8.1.2 Batch log 98

Most of the analyses in the **Toolbox** are able to perform the same analysis on several elements in one batch. This means that analyzing large amounts of data is very easily accomplished. If you e.g. wish to translate a large number of DNA sequence to protein, you can just select the DNA sequences and set the parameters for the translation once. Each DNA sequence will then be treated individually as if you performed the translation on each of them. The process will run in the background and you will be able to work on other projects at the same time.

8.1 How to handle results of analyses

All the analyses in the **Toolbox** are performed in a step-by-step procedure. First, you select elements for analyses, and then there are a number of steps where you can specify parameters (some of the analyses have no parameters, e.g. when translating DNA to RNA). The final step concerns the handling of the results of the analysis, and it is almost identical for all the analyses so we explain it in this section in general.

In this step, shown in figure 8.1, you have two options:

- **Open.** This will open the result of the analysis in a view. This is the default setting.
- **Save.** This means that the result will not be opened but saved to a folder in the **Navigation Area**. If you select this option, click **Next** and you will see one more step where you can specify where to save the results (see figure 8.2). In this step, you *have to select a folder*. You also have the option of creating a new folder in this step.

8.1.1 When the analysis does not create new elements

When an analysis does not create new elements, as e.g. **Find Open Reading Frames** which adds annotations to the sequences, the options for saving are different (see figure 8.3):

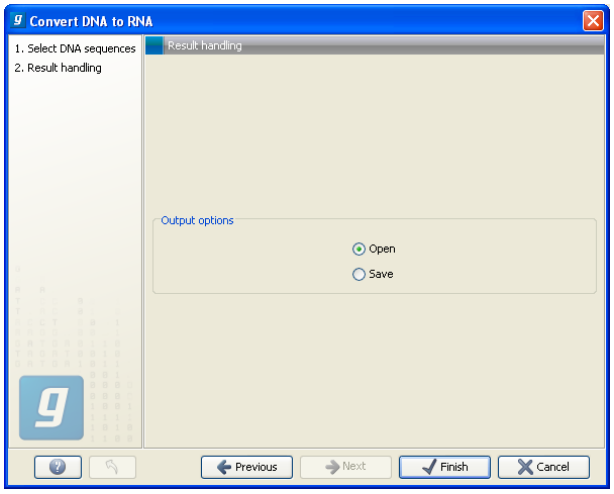


Figure 8.1: *The last step of the analyses exemplified by Translate DNA to RNA.*

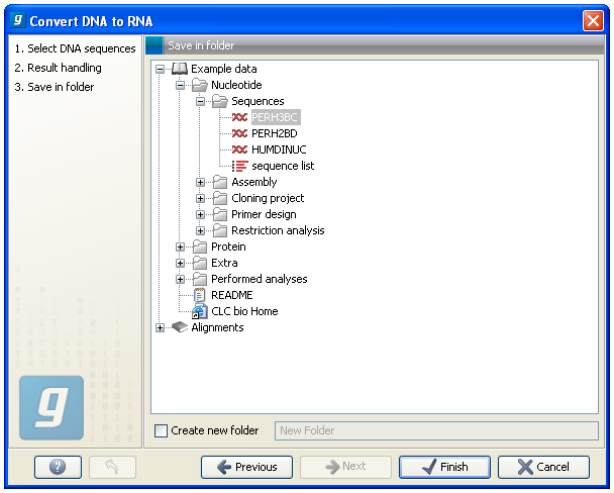


Figure 8.2: *Specify a folder for the results of the analysis.*

- **Open.** This will open each of the selected sequences in a view.
- **Save.** This will not open the sequences but just add the annotations.
- **Copy and save in new folder.** This option does not add annotations to the existing sequences but saves a copy of the selected sequences. Choosing this option means that there will be an extra step for selecting a folder where the copies of the sequences can be saved.

8.1.2 Batch log

For some analyses, there is an extra option in the final step to create a log of the batch process. This log will be created in the beginning of the process and continually updated with information about the results. See an example of a log in figure 8.4. In this example, the log displays information about how many open reading frames were found.

The log will either be saved with the results of the analysis or opened in a view with the results, depending on how you chose to handle the results.

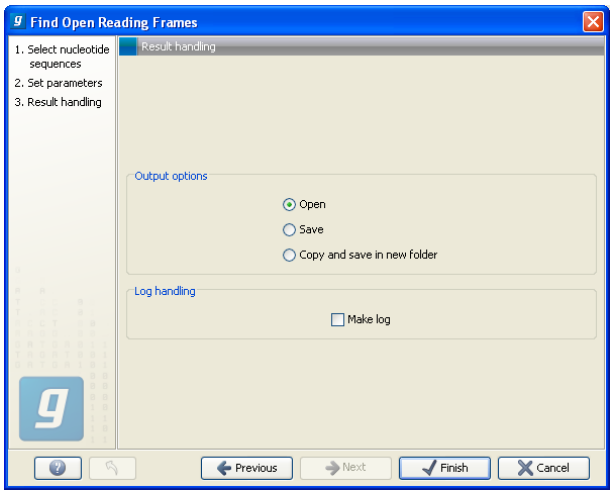


Figure 8.3: *The final step when the analysis does not create new elements but add annotations to existing elements.*

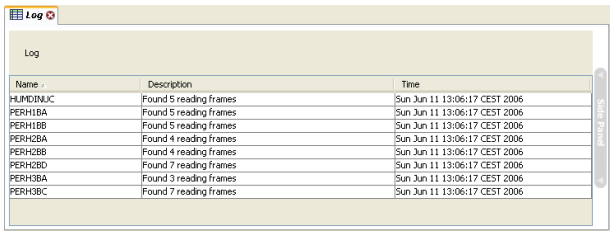

Log
Log
Name
Description
Time
HUMDINUC Found 5 reading frames Sun Jun 11 13:06:17 CEST 2006
PERH1BA Found 5 reading frames Sun Jun 11 13:06:17 CEST 2006
PERH1BB Found 5 reading frames Sun Jun 11 13:06:17 CEST 2006
PERH2BA Found 4 reading frames Sun Jun 11 13:06:17 CEST 2006
PERH2BB Found 4 reading frames Sun Jun 11 13:06:17 CEST 2006
PERH2BD Found 7 reading frames Sun Jun 11 13:06:17 CEST 2006
PERH3BA Found 3 reading frames Sun Jun 11 13:06:17 CEST 2006
PERH3BC Found 7 reading frames Sun Jun 11 13:06:17 CEST 2006

Figure 8.4: *An example of a batch log when finding open reading frames.*

Part III

Bioinformatics

Chapter 9

Database search

Contents

9.1 GenBank search 101
9.1.1 GenBank search options 101
9.1.2 Handling of GenBank search results 103
9.2 Sequence web info 104
9.2.1 Google sequence 105
9.2.2 NCBI 105
9.2.3 PubMed References 105
9.2.4 UniProt 106

CLC Gene Workbench 2.0 offers different ways of searching data on the Internet. You must be online when initiating and performing the following searches:

9.1 GenBank search

This section describes searches in GenBank - the **NCBI Entrez** database - and the import of search results. The NCBI search view is opened in this way (figure 9.1):

Search | **Search NCBI Entrez (** **)**

or **Ctrl + B (⌘ + B on Mac)**

This opens the following view:

9.1.1 GenBank search options

Conducting a search in the **NCBI Database** from *CLC Gene Workbench 2.0* corresponds to conducting the search on NCBI's website. When conducting the search from *CLC Gene Workbench 2.0*, the results are available and ready to work with straight away.

You can choose whether you want to search for nucleotide sequences or protein sequences.

As default, *CLC Gene Workbench*offers one text field where the search parameters can be entered. Click **Add search parameters** to add more parameters to your search.

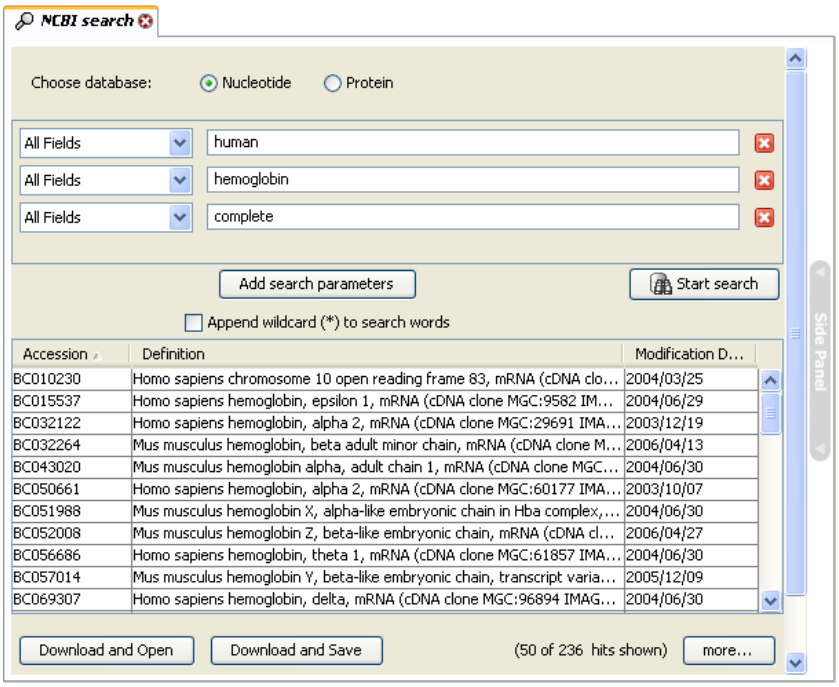

NCBI search
Choose database:
Nucleotide
Protein
All Fields
human
All Fields
hemoglobin
All Fields
complete
Add search parameters
Start search
Append wildcard (*) to search words
Accession
Definition
Modification D...
BC010230 Homo sapiens chromosome 10 open reading frame 83, mRNA (cDNA clo... 2004/03/25
BC015537 Homo sapiens hemoglobin, epsilon 1, mRNA (cDNA clone MGC:9582 IM... 2004/06/29
BC032122 Homo sapiens hemoglobin, alpha 2, mRNA (cDNA clone MGC:29691 IMA... 2003/12/19
BC032264 Mus musculus hemoglobin, beta adult minor chain, mRNA (cDNA clone M... 2006/04/13
BC043020 Mus musculus hemoglobin alpha, adult chain 1, mRNA (cDNA clone MGC... 2004/06/30
BC050661 Homo sapiens hemoglobin, alpha 2, mRNA (cDNA clone MGC:60177 IMA... 2003/10/07
BC051988 Mus musculus hemoglobin X, alpha-like embryonic chain in Hba complex,... 2004/06/30
BC052008 Mus musculus hemoglobin Z, beta-like embryonic chain, mRNA (cDNA cl... 2006/04/27
BC056686 Homo sapiens hemoglobin, theta 1, mRNA (cDNA clone MGC:61857 IMA... 2004/06/30
BC057014 Mus musculus hemoglobin Y, beta-like embryonic chain, transcript varia... 2005/12/09
BC069307 Homo sapiens hemoglobin, delta, mRNA (cDNA clone MGC:96894 IMAG... 2004/06/30
Download and Open
Download and Save
(50 of 236 hits shown)
more...
Side Panel

Figure 9.1: *The GenBank search dialog.*

Notice! The search is a "and" search, meaning that when adding search parameters to your search, you search for both (or all) text strings rather than "any" of the text strings.

You can append a wildcard character by checking the checkbox at the bottom. This means that you only have to enter the first part of the search text, e.g. searching for "genom" will find both "genomic" and "genome".

The following parameters can be added to the search:

- **All fields**. Text, searches in all parameters in the NCBI database at the same time.
- **Organism**. Text.
- **Description**. Text.
- **Modified Since**. Between 30 days and 10 years.
- **Gene Location**. Genomic DNA/RNA, Mitochondrion, or Chloroplast.
- **Molecule**. Genomic DNA/RNA, mRNA or rRNA.
- **Sequence Length**. Number for maximum or minimum length of the sequence.
- **Gene Name**. Text.

The search parameters are the most recently used. The **All fields** allows searches in all parameters in the NCBI database at the same time. **All fields** also provide an opportunity to restrict a search to parameters which are not listed in the dialog. E.g. writing 'gene[Feature key] AND mouse' in **All fields** generates hits in the GenBank database which contains one or more genes and where 'mouse' appears somewhere in GenBank file. NB: the 'Feature Key' option is only available in GenBank when searching for nucleotide sequences. For more information about

how to use this syntax, see http://www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary_Matrices.html#Search_Fields_and_Qualifiers

When you are satisfied with the parameters you have entered, you can either **Save search parameters** or **Start search**.
When applying he **Save search parameters** option, only the parameters are saved - not the results of the search. The search parameters can also be saved by dragging the tab of the Search view into the **Navigation Area**.

If you don't save the search, the search parameters are saved in **Search NCBI** view until the next time you conduct an NCBI search.

Notice! When conducting a search, no files are downloaded. Instead, the program produces a list of links to the files in the NCBI database. This ensures a much faster search.

The search process runs in the **Toolbox** under the **Processes** tab. It is possible to stop the search process by clicking stop ( ).

Because the process runs in the **Processes** tab it is possible to perform other tasks while the search is running.

9.1.2 Handling of GenBank search results

The search result is presented as a list of links to the files in the NCBI database. The **View** displays 50 hits at a time (can be changed in the **Preferences** (see chapter 4). More hits can be displayed by clicking the **More...** button at the bottom right of the **View**.

Each sequence hit is represented by text in three columns:

- Accession.
- Definition.
- Modification date.

It is possible to exclude one or more of these columns by adjust the View preferences for the database search view. Furthermore, your changes in the View preferences can be saved. See section 4.5.

Several sequences can be selected, and by clicking the buttons in the bottom of the search view, you can do the following:

- Download and open, doesn't save the sequence.
- Download and save, lets you choose location for saving sequence.
- Open at NCBI, searches the sequence at NCBI's web page.

Double-clicking a hit will download and open the sequence. The hits can also be copied into the **View Area** or the **Navigation Area** from the search results by drag and drop, copy/paste or by using the right-click menu. Finally, you can also

Drag and drop from GenBank search results

The sequences from the search results can be opened by dragging them into a position in the **View Area**.

Notice! A sequence is not saved until the **View** displaying the sequence is closed. When that happens, a dialog opens: Save changes of sequence x? (Yes or No).

The sequence can also be saved by dragging it into the **Navigation Area**. It is possible to select more sequences and drag all of them into the **Navigation Area** at the same time.

Download GenBank search results using right-click menu

You may also select one or more sequences from the list and download using the right-click menu (see figure 9.2). Choosing **Save sequence** lets you select a folder or project where the sequences are saved when they are downloaded. Choosing **Open sequence** opens a new view for each of the selected sequences.

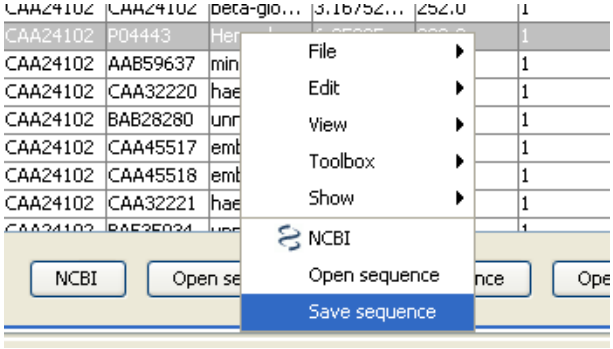


Figure 9.2: *By right-clicking a search result, it is possible to choose how to handle the relevant sequence.*

Copy/paste from GenBank search results

When using copy/paste to bring the search results into the **Navigation Area**, the actual files are downloaded from GenBank.

To copy/paste files into the **Navigation Area**:

> **select one or more of the search results | Ctrl + C (⌘ + C on Mac) | select project or folder in the Navigation Area | Ctrl + V**

Notice! Search results are downloaded before they are saved. Downloading and saving several files may take some time. However, since the process runs in the background (displayed in the **Status bar**) it is possible to continue other tasks in the program. Like the search process, the download process can be stopped. This is done in the **Toolbox** in the **Processes** tab.

9.2 Sequence web info

CLC Gene Workbench 2.0 provides direct access to web-based search in various databases and on the Internet using your computer's default browser. You can look up a sequence in the databases of NCBI and UniProt, search for a sequence on the Internet using Google and

search for Pubmed references at NCBI. This is useful for quickly obtaining updated and additional information about a sequence.

The functionality of these search functions depends on the information that the sequence contains. You can see this information by viewing the sequence as text (see section 11.3). In the following sections, we will explain this in further detail.

The procedure for searching is identical for all four search options (see also figure 9.3):

Right-click a sequence in the Navigation Area | Sequence Web Info | select the desired search function

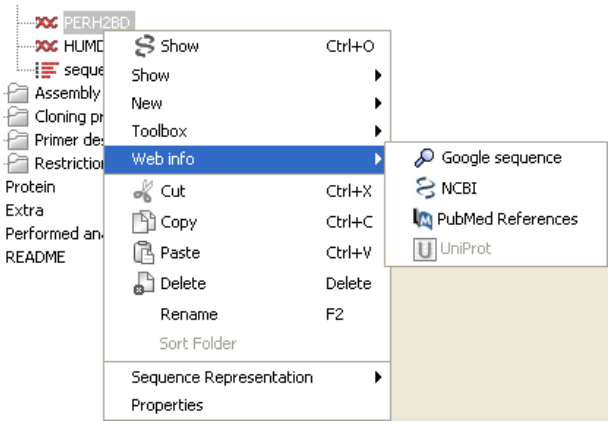


Figure 9.3: *By right-clicking a search result, it is possible to choose how to handle the relevant sequence.*

This will open your computer's default browser searching for the sequence that you selected.

9.2.1 Google sequence

The Google search function uses the accession number of the sequence which is used as search term on `http://www.google.com`. The resulting web page is equivalent to typing the accession number of the sequence into the search field on `http://www.google.com`.

9.2.2 NCBI

The NCBI search function searches in GenBank at NCBI (`http://www.ncbi.nlm.nih.gov`) using an identification number (when you view the sequence as text it is the "GI" number). Therefor, the sequence file must contain this number in order to look it up in NCBI. All sequences downloaded from NCBI have this number.

9.2.3 PubMed References

The PubMed references search option lets you look up Pubmed articles based on references contained in the sequence file (when you view the sequence as text it contains a number of "PUBMED" lines). Not all sequence have these PubMed references, but in this case you will se a dialog and the browser will not open.

9.2.4 UniProt

The UniProt search function searches in the UniProt database (http://www.ebi.uniprot.org) using the accession number. Furthermore, it checks whether the sequence was indeed downloaded from UniProt.

Chapter 10

BLAST Search

Contents

10.1 BLAST Against NCBI Database 107
10.1.1 Output from BLAST search 110
10.1.2 BLAST table 112
10.2 BLAST Against Local Database 113
10.3 Create Local BLAST Database 114

CLC Gene Workbench offers to conduct BLAST searches on protein and DNA sequences. In short, a BLAST search identifies homologous sequences by searching one or more databases hosted by NCBI (`http://www.ncbi.nlm.nih.gov/`), on your query sequence [McGinnis and Madden, 2004]. BLAST (Basic Local Alignment Search Tool), identifies homologous sequences using a heuristic method which finds short matches between two sequences. After initial match BLAST attempts to start local alignments from these initial matches.

From *CLC Gene Workbench 2.0* it is also possible to conduct BLAST searches on a database stored locally on your computer. Local BLAST and the creation of a database for local BLAST search is described later in this chapter.

10.1 BLAST Against NCBI Database

To conduct a BLAST search:

right-click the tab of an open sequence | **Toolbox** | **BLAST Search( )** | **BLAST Against NCBI Databases ( )**

or **click an element in the Navigation Area** | **Toolbox** | **BLAST Search( )** | **BLAST Against NCBI Databases ( )**

Alternatively, use the keyboard shortcut: Ctrl+Shift+B for Windows and ⌘ +Shift+B on Mac OS. This opens the BLAST dialog. You can not use sequences longer than 8190 for BLAST search.

This opens the dialog seen in figure 10.1

Click **Next**

In **Step 2**, you can choose which type of BLAST search you want to conduct, and you can limit your

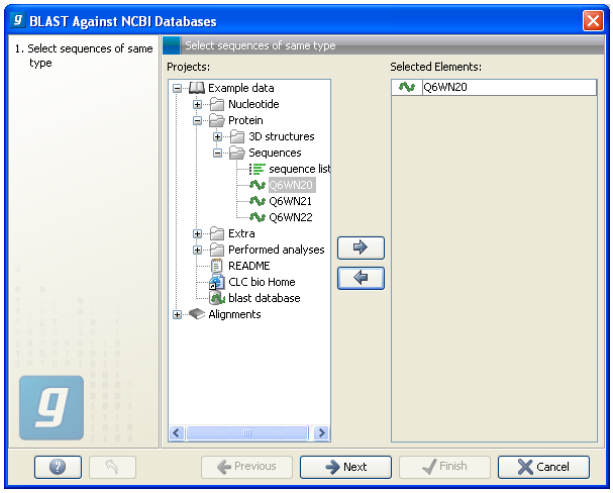


Figure 10.1: *Choose one or more sequences to conduct a BLAST search.*

search to a particular database (see section B in the appendix for a list of available databases). Step 2 can be seen in figure 10.2:

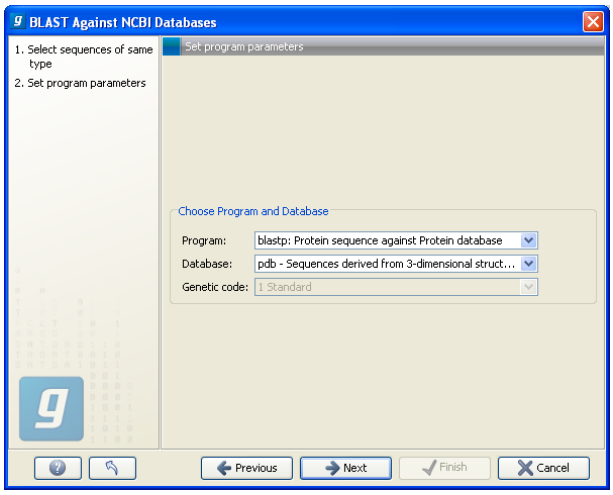


Figure 10.2: *Choose a BLAST Program and a database for the search.*

BLAST search for DNA sequences:

- **BLASTn: DNA sequence against DNA database** This BLAST method is used to identify homologous DNA sequences to your query sequence.
- **BLASTx: Translated DNA sequence against Protein database** If you want to search in protein databases, this BLAST method allows for automated translation of the DNA input sequence and searching in various protein databases.
- **tBLASTx: Translated DNA sequence against Translated DNA database** Here is both the input DNA sequence and the searched DNA database automatically translated.

BLAST search for protein sequences:

- **BLASTp: Protein sequence against Protein database** This the most common BLAST method used when searching for homologous protein sequences having a protein sequence as search input.

- **tBLASTn: Protein sequence against Translated DNA database** Here is the protein sequence searched against an automatically translated DNA database.

Depending on whether you choose a protein or a DNA sequence, a number of different databases can be searched. A complete list of these databases can be found in Appendix B. When **nr** appears in the **Database** parameter drop down menu, the search will include all relevant databases at NCBI. The **nr** database is the most complete, but also the most redundant database that can be searched. Searches can be limited to less complete databases. As an example, when choosing **pdb** only sequences with a know structure are searched. If homologous sequences are found to the query sequence, these can be downloaded and opened with the 3D viewer of *CLC Protein Workbench*.

When choosing BLASTx or tBLASTx to conduct a search, you get the option of selecting a translation table for the genetic code. The standard genetic code is set as default. This is particularly useful when working with organisms or organelles which have a genetic code that differs from the standard genetic code.

In **Step 3** you can limit the BLAST search by adjusting the parameters seen in figure 10.3

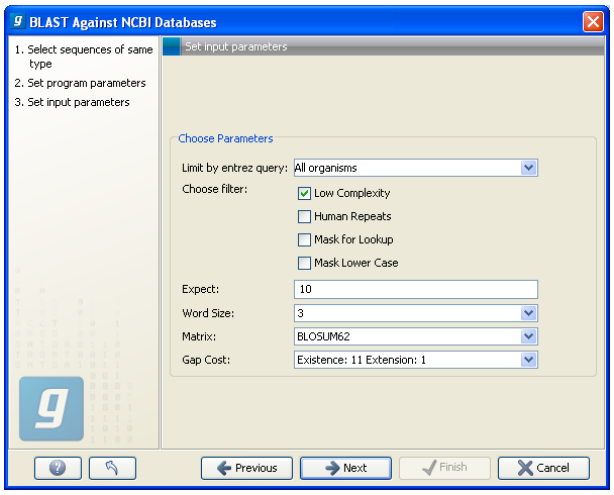


Figure 10.3: *Examples of different limitations which can be set before submitting a BLAST search.*

The following description of BLAST search parameters is based on information from `http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml`.

- **Limit by Entrez query** BLAST searches can be limited to the results of an Entrez query against the database chosen. This can be used to limit searches to subsets of the BLAST databases. Any terms can be entered that would normally be allowed in an Entrez search session. Some queries are preentered and can be chosen in the drop down menu.
- **Choose filter**
  - Low-complexity. Mask off segments of the query sequence that have low compositional complexity. Filtering can eliminate statistically significant, but biologically uninteresting reports from the BLAST output (e.g. hits against common acidic-, basic- or proline-rich regions), leaving the more biologically interesting regions of the query sequence available for specific matching against database sequences.

  - Human Repeats. This option masks Human repeats (LINE's and SINE's) and is especially useful for human sequences that may contain these repeats. Filtering for repeats can increase the speed of a search especially with very long sequences (>100 kb) and against databases which contain large number of repeats (htgs).
  - Mask for Lookup. This option masks only for purposes of constructing the lookup table used by BLAST. BLAST searches consist of two phases, finding hits based upon a lookup table and then extending them.
  - Mask Lower Case. With this option selected you can cut and paste a FASTA sequence in upper case characters and denote areas you would like filtered with lower case. This allows you to customize what is filtered from the sequence during the comparison to the BLAST databases

- **Expect** The statistical significance threshold for reporting matches against database sequences: the default value is 10, meaning that 10 matches are expected to be found merely by chance, according to the stochastic model of Karlin and Altschul (1990). If the statistical significance ascribed to a match is greater than the EXPECT threshold, the match will not be reported. Lower EXPECT thresholds are more stringent, leading to fewer chance matches being reported. Increasing the threshold shows less stringent matches. Fractional values are acceptable.

- **Word Size** BLAST is a heuristic that works by finding word-matches between the query and database sequences. You may think of this process as finding "hot-spots" that BLAST can then use to initiate extensions that might lead to full-blown alignments. For nucleotide-nucleotide searches (i.e. "BLASTn") an exact match of the entire word is required before an extension is initiated, so that you normally regulate the sensitivity and speed of the search by increasing or decreasing the wordsize. For other BLAST searches non-exact word matches are taken into account based upon the similarity between words. The amount of similarity can be varied so that you normally uses just the wordsizes 2 and 3 for these searches.

- **Matrix** A key element in evaluating the quality of a pairwise sequence alignment is the "substitution matrix", which assigns a score for aligning any possible pair of residues. The matrix used in a BLAST search can be changed depending on the type of sequences you are searching with (see the BLAST Frequently Asked Questions).

- **Gap Cost** The pull down menu shows the Gap Costs (Penalty to open Gap and penalty to extend Gap). There can only be a limited number of options for these parameters. Increasing the Gap Costs and Lambda ratio will result in alignments which decrease the number of Gaps introduced.

The more limitations are submitted to the search parameters, the faster the search will be conducted. If no limitations are submitted, the BLAST search may take several minutes.

When the **Advanced parameters** of **Step 3** are adjusted, click **Next** to choose whether you want to open the BLAST output in an editor and/or in a table. 10.1.1

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

10.1.1 Output from BLAST search

The two different outputs from a BLAST search are shown in figure 10.4:

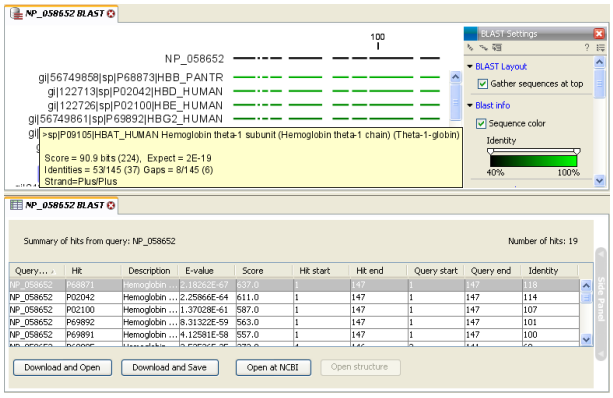

NP_058652 BLAST
100
NP_058652
gi|56749858|sp|P68873|HBB_PANTR
gi|122713|sp|P02042|HBD_HUMAN
gi|122726|sp|P02100|HBE_HUMAN
gi|56749861|sp|P69892|HBG2_HUMAN
>sp|P09105|HBAT_HUMAN Hemoglobin theta-1 subunit (Hemoglobin theta-1 chain) (Theta-1-globin)
Score = 90.9 bits (224), Expect = 2E-19
Identities = 53/145 (37) Gaps = 8/145 (6)
Strand=Plus/Plus
BLAST Settings
BLAST Layout
Gather sequences at top
Blast info
Sequence color
Identity
40%
100%
NP_058652 BLAST
Summary of hits from query: NP_058652
Number of hits: 19
Query... Hit Description E-value Score Hit start Hit end Query start Query end Identity
NP_058652 P68871 Hemoglobin ... 2.18262E-67 637.0 1 147 1 147 118
NP_058652 P02042 Hemoglobin ... 2.25866E-64 611.0 1 147 1 147 114
NP_058652 P02100 Hemoglobin ... 1.37028E-61 587.0 1 147 1 147 107
NP_058652 P69892 Hemoglobin ... 8.31322E-59 563.0 1 147 1 147 101
NP_058652 P69891 Hemoglobin ... 4.12581E-58 557.0 1 147 1 147 100
Download and Open
Download and Save
Open at NCBI
Open structure

Figure 10.4: *Display of the output of a BLAST search. At the top is there a graphical representation of BLAST hits with tool-tips showing additional information on individual hits. Below is shown a tabular form of the BLAST results.*

The **BLAST Graphics** and the **BLAST table** are described in the following chapters.

BLAST Graphics

The **BLAST editor** shows the sequences hits which were found in the BLAST search. The hit sequences are represented by colored horizontal lines, and when hovering the mouse pointer over a BLAST hit sequence, a tooltip appears, listing the characteristics of the sequence.

There are several View preferences available for in the **BLAST Graphics** view.

- **BLAST Layout** You can choose whether to **Gather sequences at top**, which means that vertical gaps between sequences are eliminated to assist comparison between the query sequence and the hit sequences.
- **BLAST info** In this View preference group, you can choose whether to color hit sequences and you can adjust the coloring.

The remaining View preferences for BLAST Graphics are the same as those of alignments. See section 18.2.

Some of the information available in the tooltips is:

- **Name of sequence** Here is shown some additional information of the sequence which was found. This line corresponds to the description line in GenBank (if the search was conducted on the nr database).
- **Length of sequence** This shows the entire length of the found sequence.
- **Score** This shows the bit score of the local alignment generated through the BLAST search.
- **Expect** Also known as the E-value. A low value indicates a homologous sequence. Higher E-values indicate that BLAST found a less homologous sequence.
- **Identities** This number shows the number of identical residues or nucleotides in the obtained alignment.
- **Gaps** This number shows whether the alignment has gaps or not.

- **Strand** This is only valid for nucleotide sequences and show the direction of the aligned strands. Minus indicate a complementary strand.
- **Query** This is the sequence (or part of the sequence) which you have used for the BLAST search.
- **Sbjct (subject)** This is the sequence found in the database.

The numbers of the query and subject sequences refer to the sequence positions in the submitted and found sequences. If the subject sequence has number 59 in front of the sequence, this means that 58 residues are found upstream of this position, but these are not included in the alignment.

In addition to the latter described output of a BLAST search, it is possible to view the BLAST results in a tabular view. In the tabular view, one can get a quick and fast overview of the results. In the tabular view it is possible to select multiple sequences and for example download all of these in one single step. Moreover, it is possible to look additional information on each single hit is the BLAST result on the NCBI homepage. These possibilities are either available through a right-click with the mouse or by using the buttons at the end of the table.

10.1.2 BLAST table

If the **BLAST table** view was not selected in **Step 4** of the BLAST search, the table can be generated in the following way:

Right-click the tab of the initial BLAST result view | Show | BLAST Table

Figure 10.5 is an example of a BLAST Table.

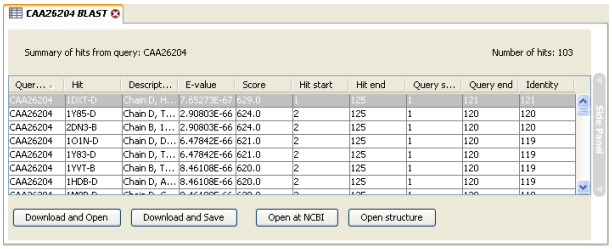

CAA26204 BLAST
Summary of hits from query: CAA26204
Number of hits: 103
Quer...
Hit
Descript...
E-value
Score
Hit start
Hit end
Query s...
Query end
Identity
CAA26204 1DXT-D Chain D, H... 7.65273E-67 629.0 1 125 1 121 121
CAA26204 1Y85-D Chain D, T... 2.90803E-66 624.0 2 125 1 120 120
CAA26204 2DN3-B Chain B, 1... 2.90803E-66 624.0 2 125 1 120 120
CAA26204 1O1N-D Chain D, D... 6.47842E-66 621.0 2 125 1 120 119
CAA26204 1Y83-D Chain D, T... 6.47842E-66 621.0 2 125 1 120 119
CAA26204 1YVT-B Chain B, T... 8.46108E-66 620.0 2 125 1 120 119
CAA26204 1HDB-D Chain D, A... 8.46108E-66 620.0 2 125 1 120 119
Download and Open
Download and Save
Open at NCBI
Open structure

Figure 10.5: *Display of the output of a BLAST search in the tabular view. The hits can be sorted by the different columns, simply by clicking the column heading.*

The BLAST Table includes the following information:

- **Query sequence** The sequence which was used for the search.
- **Hit** The Name of the sequences found in the BLAST search
- **Description** Text from NCBI describing the sequence.
- **E-value** Measure of quality of the match. Higher E-values indicate that BLAST found a less homologous sequence.
- **Score** This shows the bit score of the local alignment generated through the BLAST search.
- **Hit start** Shows the start position in the hit sequence

- **Hit end** Shows the end position in the hit sequence
- **Query start** Shows the start position in the query sequence
- **Query end** Shows the end position in the query sequence
- **Identity** Shows the number of identical residues in the query and hit sequence

In the **BLAST table** view you can handle the hit sequences. Select one or more sequences from the table, and apply one of the following functions.

- **NCBI**
  Opens the corresponding sequence(s) at GenBank at NCBI. Here is stored additional information regarding the selected sequence(s). The default Internet browser is used for this purpose.
- **Open sequence**
  Opens the selected sequence(s) in one or more sequence views.
- **Save sequence**
  Downloads and saves the sequence without opening it.
- **Open structure** If the hit sequence contain structure information, the sequence is opened in a text view or a 3D view (3D view in CLC Protein Workbench and CLC Combined Workbench).

10.2 BLAST Against Local Database

CLC Gene Workbench will let you conduct a BLAST search in a local database. See section 10.3 for more about how to create a database.

The advantage of conducting a local BLAST search is the speed and that it is possible to BLAST sequences longer than 8900 residues.

To conduct a Local BLAST search:

right-click the tab of an open sequence | **Toolbox** | **BLAST Search()** | **BLAST Against Local Databases ( )**

or **click an element in the Navigation Area** | **Toolbox** | **BLAST Search( )** | **BLAST Against Local Databases ( )**

This opens the dialog seen in figure 10.6:

Click **Next**

This opens the dialog seen in figure 10.7:

In **Step 2**,you can choose between different BLAST methods. See section 10.1 for information about these methods.

In **step 2** you can also choose which of your local BLAST databases you want to conduct the search in. Clicking **Select Database** opens the dialog shown in figure 10.8:

Select a Click **Next**

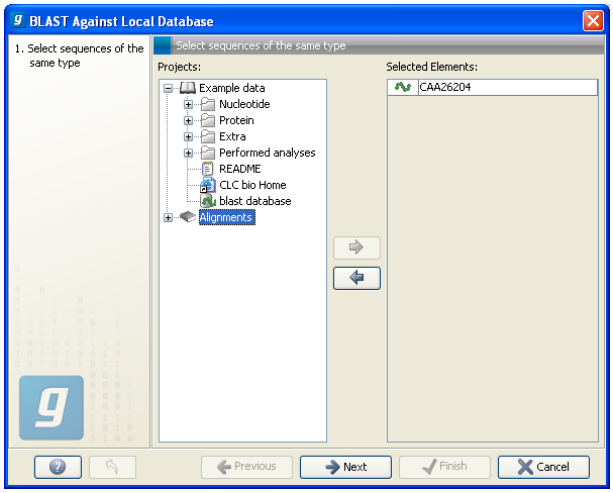


Figure 10.6: *Choose one or more sequences to conduct a Local BLAST search.*

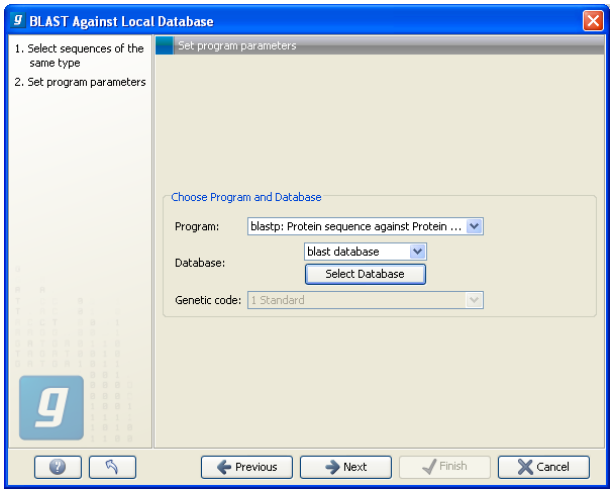


Figure 10.7: *Choose a BLAST program and a local database to conduct BLAST search.*

This opens the dialog seen in figure 10.9:

See section 10.1 for information about these limitations.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

10.3 Create Local BLAST Database

In *CLC Gene Workbench* you can create a local database which you can use for local BLAST. Both DNA, RNA, and protein sequences can be used.

It is not necessary to import the sequences into *CLC Gene Workbench* before creating the database. The local database can be created from sequences which are stored in the **Navigation Area** or the sequences can be browsed from the computer's file system. In the latter case, the files must be in fasta (.fsa/.fa/.fasta) format.

To create a local BLAST data base from the file system or from the **Navigation Area**:

BLAST search in Toolbox( ) | Create Local BLAST Database( )

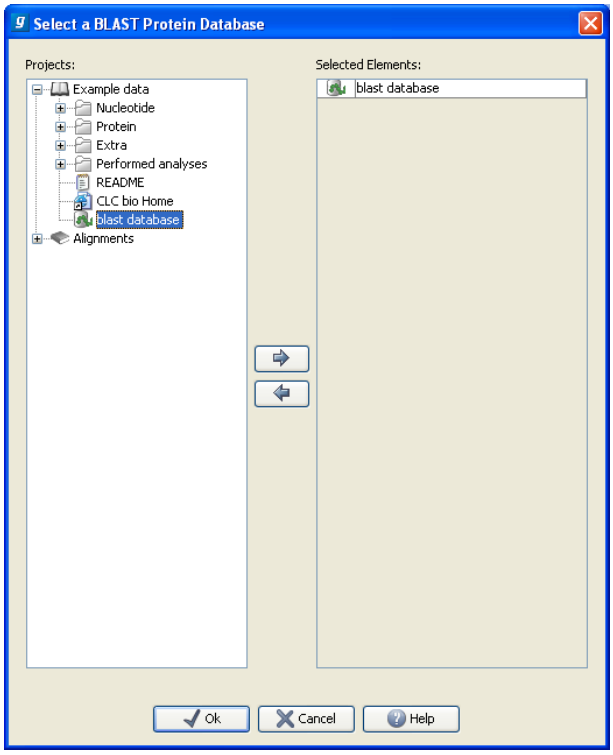


Figure 10.8: *Select your local BLAST database.*

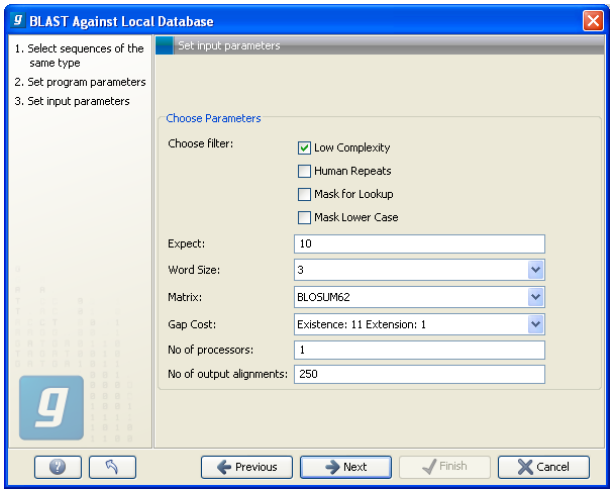


Figure 10.9: *Examples of different limitations which can be set before submitting a BLAST search.*

This opens the dialog seen in figure 10.10

- **Select Input Source.** Lets you choose whether to include sequences from the **Navigation Area** or from the computer's file system (External FASTA file).
- **Sequence type.** If you choose to import sequences from an external FASTA file into the database, you must choose whether the sequences are nucleotide or protein sequences.
- **FormatDB Option.** Enables or disables parsing of SeqId and creation of indeces.
- **Input Sequences.** Depending on the choice of **Select Input Source** above, clicking the button will let you browse the **Navigation Area** or the external file system for the sequences

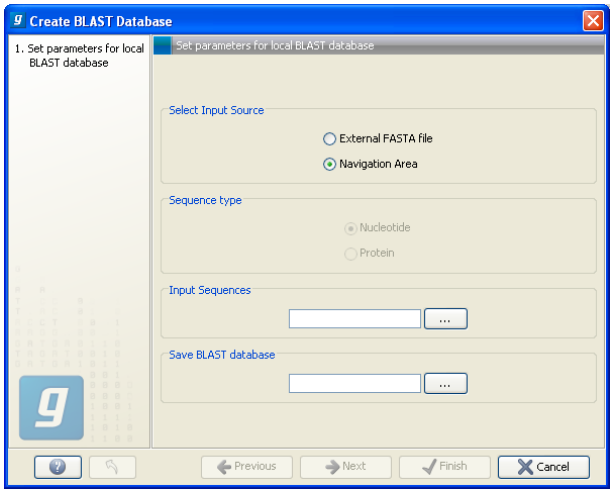


Figure 10.10: *Setting parameters for the local BLAST database.*

which you want to include in the database.

- **Save BLAST database.** Lets you browse your external file system for a suitable place to save the database.

After having adjusted all these settings, click **Next**, which opens the dialog seen in figure 10.11

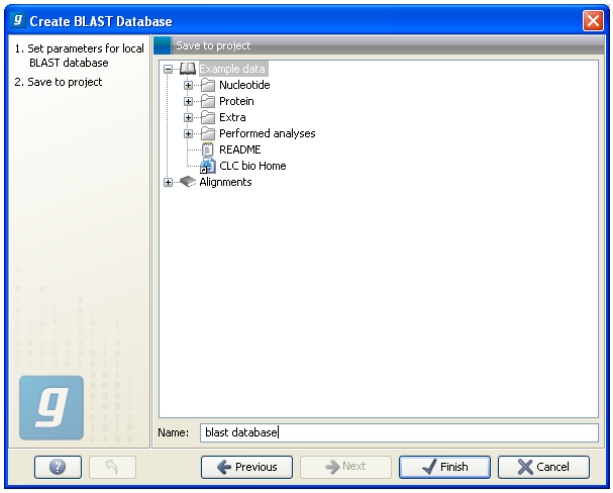


Figure 10.11: *Choose where the access point to your local BLAST database is saved in the Navigation Area.*

Click **Next** to complete the creation of the database.

Chapter 11

Viewing and editing sequences

Contents

11.1 View sequence . . . 117
11.1.1 Sequence Layout in Side Panel . . . 118
11.1.2 Selecting parts of the sequence . . . 123
11.1.3 Editing the sequence . . . 124
11.1.4 Adding and modifying annotations . . . 124
11.1.5 Removing annotations . . . 126
11.1.6 Sequence region types . . . 126
11.2 Sequence information . . . 127
11.2.1 Annotation map . . . 128
11.3 View as text . . . 128
11.4 Creating a new sequence . . . 129
11.5 Sequence Lists . . . 130
11.5.1 Graphical view of sequence lists . . . 131
11.5.2 Sequence list table . . . 132
11.5.3 Extract sequences . . . 132
11.6 Circular DNA . . . 132
11.6.1 Show restriction sites for circular DNA . . . 133
11.6.2 Using split views to see details of the circular molecule . . . 134
11.6.3 Mark molecule as circular and specify starting point . . . 134

CLC Gene Workbench 2.0 offers three different ways of viewing and editing sequences as described in this chapter. Furthermore, this chapter also explains how to create a new sequence and how to assemble several sequences in a sequence list.

11.1 View sequence

When you double-click a sequence in the **Navigation Area**, the sequence will open automatically, and you will see the nucleotides or amino acids. The zoom options described in section 3.3 allow you to e.g. zoom out in order to see more of the sequence in one view. There are a number of options for viewing and editing the sequence which are all described in this section. All the options described in this section also apply to alignments (further described in section 18.2).

11.1.1 Sequence Layout in Side Panel

Each view of a sequence has a **Side Panel** located at the right side of the view. When you make changes in the **Side Panel** the view of the sequence is instantly updated. To show or hide the **Side Panel**:

select the View | Ctrl + U

or **Click the ( ) at the top right corner of the Side Panel to hide | Click the gray Side Panel button to the right to show**

When you open a view, the **Side Panel** has default settings which can be changed in the **User Preferences** (see chapter 4).

Below, each group of preferences will be explained. Some of the preferences are not the same for nucleotide and protein sequences, but the differences will be explained for each group of preferences.

Notice! When you make changes to the settings in the **Side Panel**, they are not automatically saved when you save the sequence. Click **Save/restore Settings** ( ) to save the settings (see section 4.5 for more information).

Sequence Layout

These preferences determine the overall layout of the sequence:

- **Space every 10 residues.** Inserts a space every 10 residues - only visible when you zoom in to see the residues.
- **Wrap sequences.** Shows the sequence on more than one line.
  - **No wrap.** The sequence is displayed on one line.
  - **Auto wrap.** Wraps the sequence to fit the width of the view, not matter if it is zoomed in our out (displays minimum 10 nucleotides on each line).
  - **Fixed wrap.** Makes it possible to specify when the sequence should be wrapped. In the text field below, you can choose the number of residues to display on each line.
- **Double stranded.** Shows both strands of a sequence (only applies to DNA sequences).
- **Numbers on plus strand.** Whether to set the numbers relative to the positive or the negative strand in a nucleotide sequence – (only applies to DNA sequences).
- **Numbers on sequences.** Shows residue positions along the sequence. The starting point can be changed by setting the number in the field below. If you set it to e.g. 101, the first residue will have the position of -100. This can also be done by right-clicking an annotation and choosing **Set Numbers Relative to This Annotation**.
- **Follow selection.** When viewing the same sequence in two separate views, "Follow selection" will automatically scroll the view in order to follow a selection made in the other view.
- **Lock numbers.** When you scroll vertically, the position numbers remain visible. (Only possible when the sequence is not wrapped.)

- **Lock labels.** When you scroll horizontally, the label of the sequence remains visible.
- **Sequence label.** Defines the label to the left of the sequence.
  - Name (this is the default information to be shown).
  - Accession (sequences downloaded from databases like GenBank have an accession number).
  - Species.
  - Species (accession).
  - Common Species.
  - Common Species (accession).

Annotation Layout

Annotations are data attached to a specific part of a sequence. If the sequence is downloaded from a database it has annotations attached to it, e.g. the location of genes on a DNA sequence. If you have performed **Restriction Site** or **Proteolytic Cleavage** analysis, the cut sites can be displayed as annotations on the sequence. Other analyses also attach annotations on the sequence. See section 11.1.6 for more information about how to interpret the annotations. The annotations are shown as colored boxes along the sequence, and their appearance is determined in the **Annotation layout** preferences group:

- **Show annotations.** Determines whether the annotations are shown.
- **Position.**
  - **On sequence.** The annotations are placed on the sequence. The residues are visible through the annotations (if you have zoomed in to 100%).
  - **Next to sequence.** The annotations are placed above the sequence.
- **Offset.** If several annotations cover the same part of a sequence, they can be spread out.
  - **Piled.** The annotations are piled on top of each other. Only the one at front is visible.
  - **Little offset.** The annotations are piled on top of each other, but they have been offset a little.
  - **More offset.** Same as above, but with more spreading.
  - **Most offset.** The annotations are placed above each other with a little space between. This can take up a lot of space on the screen.
- **Label.** Each annotation can be labelled with a name. Additional information about the sequence is shown if you place the mouse cursor on the annotation and keep it still.
  - **No labels.** No labels are displayed.
  - **On annotation.** The labels are displayed in the annotation's box.
  - **Over annotation.** The labels are diplayed above the annotations.
  - **Before annotation.** The labels are placed just to the left of the annotation.
  - **Flag.** The labels are displayed as flags at the beginning of the annotation.
- **Show arrows.** Toggles the display of arrow heads on the annotations.
- **Use gradients.** Fills the boxes with gradient color.

Annotation types

- **Annotation types.** This group lists all the types of annotations that are attached to the sequence that is viewed. For sequences with many annotations it can be easier to get an overview, if you deselect the annotation types that are not relevant. If you want to remove single annotations while preserving other annotations of the same type, see section 11.1.4.

It is possible to color the different annotations for better overview.

Color settings for an annotation can be done by clicking the colored square next to the relevant annotation type.

Many different settings can be set in the three layers: Swatches, HSB, and RGB. Apply your settings and click **OK**. When you click **OK**, the color settings cannot be reset. The **Reset** function only works for changes made before pressing **OK**.

Restriction sites

These preferences allow you to display restriction sites on the sequence. There is a list of enzymes which are represented by different colors. By selecting or deselecting the enzymes in the list, you can specify which enzymes' restriction sites should be displayed (see figure 17.4).

Figure 11.1: *Showing restriction sites of two restriction enzymes.*

The color of the flag of the restriction site can be changed by clicking the colored box next to the enzyme's name.

The list of restriction enzymes contains per default ten of the most popular enzymes, but you can easily modify this list and add more enzymes. You have four ways of modifying the list:

- **Edit enzymes button.** This displays a dialog with the enzymes currently in the list shown at the bottom and a list of available enzymes at the top. To add more enzymes, select them in the upper list and press the **Add enzymes button** (). To remove enzymes, select them in the list below and click the **Remove enzymes button** ().
- **Load enzymes button.** If you have previously created an enzyme list, you can select this list by clicking the Load enzymes button. You can filter the enzymes in the same way as illustrated in figure 17.13.
- **Add enzymes cutting the selection to panel.** If you make a selection on the sequence, right-click, you find this option for adding enzymes. Based on the entire list of available enzymes, the enzymes cutting in the region you selected will be added to the list in the **Side Panel**.
- **Insert restriction site before/after selection.** If you make a selection on the sequence, right-click, you find this option for inserting a restriction site before or after the region you

selected. A dialog is shown where you can select an enzyme whose recognition sequence is inserted. If it was not already present in the list in the **Side Panel**, the enzyme will now be added and selected.

Finally, if you have selected a set of enzymes that you wish to keep for later use, you can click **Save enzymes** and the selected enzymes will be saved to en enzyme list. This list can then be used both when finding restriction sites from the **Toolbox** or when viewing another sequence.

Residue coloring

These preferences make it possible to color both the residue letter and set a background color for the residue.

- **Non-standard residues.** For nucleotide sequences this will color the residues that are not C, G, A, T or U. For amino acids only B, Z, and X are colored as non-standard residues.
  - **Foreground color.** Sets the color of the letter. Click the color box to change the color.
  - **Background color.** Sets the background color of the residues. Click the color box to change the color.
- **Rasmol colors.** Colors the residues according to the Rasmol color scheme.
  See `http://www.openrasmol.org/doc/rasmol.html`
  - **Foreground color.** Sets the color of the letter. Click the color box to change the color.
  - **Background color.** Sets the background color of the residues. Click the color box to change the color.
- **Polarity colors (only protein).** Colors the residues according to the polarity of amino acids.
  - **Foreground color.** Sets the color of the letter. Click the color box to change the color.
  - **Background color.** Sets the background color of the residues. Click the color box to change the color.

Nucleotide info

These preferences only apply to nucleotide sequences.

- **Translation.** Displays a translation into protein just below the nucleotide sequence. Depending on the zoom level, the amino acids are displayed with three letters or one letter.
  - **Frame.** Determines where to start the translation.
    - **+1 to -1.** Select one of the six reading frames.
    - **Selection.** This option will only take effect when you make a selection on the sequence. The translation will start from the first nucleotide selected. Making a new selection will automatically display the corresponding translation. Read more about selecting in section 11.1.2.
    - **All.** Select all reading frames at once. The translations will be displayed on top of each other.

  - **Table.** The translation table to use in the translation. For more about translation tables, see section 13.4.
  - **Only AUG start codons.** For most genetic codes, a number of codons can be start codons. Selecting this option only colors the AUG codons green.

- **Trace data.** See section 16.1.

- **G/C content.** Calculates the G/C content of a part of the sequence and shows it as a gradient of colors or as a graph below the sequence.
  - **Window length.** Determines the length of the part of the sequence to calculate. A window length of 9 will calculate the G/C content for the nucleotide in question plus the 4 nucleotides to the left and the 4 nucleotides to the right. A narrow window will focus on small fluctuations in the G/C content level, whereas a wider window will show fluctuations between larger parts of the sequence.
  - **Foreground color.** Colors the letter using a gradient, where the left side color is used for low levels of G/C content and the right side color is used for high levels of G/C content. The sliders just above the gradient color box can be dragged to highlight relevant levels of G/C content. The colors can be changed by clicking the box. This will show a list of gradients to choose from.
  - **Background color.** Sets a background color of the residues using a gradient in the same way as described above.
  - **Graph.** The G/C content level is displayed on a graph.
    * **Height.** Specifies the height of the graph.
    * **Type.** The graph can be displayed as Line plot, Bar plot or as a Color bar.
    * **Color box.** For Line and Bar plots, the color of the plot can be set by clicking the color box. For Colors, the color box is replaced by a gradient color box as described under Foreground color.

Hydrophobicity info

These preferences only apply to proteins and are described in section 14.2.2.

Search

The Search group is not a preferences group, but can be used for searching the sequence. Clicking the search button will search for the first occurrence of the search string. Clicking the search button again will find the next occurrence and so on. If the search string is found, the corresponding part of the sequence will be selected.

- **Search term.** Enter the text to search for. The search function does not discriminate between lower and upper case characters.

- **Sequence search.** Search the nucleotides or amino acids. For nucleotides, all the standard IUPAC codes can be used, e.g. RT will find both GT and AT. RT will also find e.g. AN. The IUPAC codes are available from the **Help** menu under Background Information. For amino acids, the single letter abbreviations should be used for searching. Accordingly, N (for nucleotides) and X (for proteins) can be used as a wildcard character.

- **Annotation search.** Searches the annotations on the sequence. The search is performed both on the labels of the annotations, but also on the text appearing in the tooltip that you see when you keep the mouse cursor fixed. If the search term is found, the part of the sequence corresponding to the matching annotation is selected.
- **Position search.** Finds a specific position on the sequence. In order to find an interval, e.g. from position 500 to 570, enter "500..570" in the search field. This will make a selection from position 500 to 570 (both included). Notice the two periods (..) between the start an end number.
- **Include negative strand.** When searching the sequence for nucleotides or amino acids, you can search on both strands.

This concludes the description of the **View Preferences**. Next, the options for selecting and editing sequences are described.

Text format

These preferences allow you to adjust the format of all the text in the view (both residue letters, sequence label and translations if relevant).

- **Text size.** Five different sizes.
- **Font.** Shows a list of Fonts available on your computer.
- **Bold residues.** Makes the residues bold.

11.1.2 Selecting parts of the sequence

You can select parts of a sequence:

> **Click Selection ( ) in Toolbar | Press and hold down the mouse button on the sequence where you want the selection to start | move the mouse to the end of the selection while holding the button | release the mouse button**

Alternatively, you can search for a specific interval using the search function described above.

You can select several parts of sequence by holding down the **Ctrl** button while making selections. Holding down the **Shift** button lets you extend or reduce an existing selection to the position you clicked.

If you have made a selection, you can expand it by using **Shift** and **Ctrl** keys or by using the right-click menu:

> **right-click the selection | Expand Selection | Select the number of residues to expand the selection to both sides**

To select the entire sequence:

> **right-click the sequence label to the left**

To select a part of a sequence covered by an annotation:

> **right-click the annotation | Select annotation**

A selection can be opened in a new view and saved as a new sequence:

right-click the selection | **Open selection in new view**

This opens the annotated part of the sequence in a new view. The new sequence can be saved by dragging the tab of the sequence view into the **Navigation Area**.

The process described above is also the way to manually translate coding parts of sequences (CDS) into protein. You simply translate the new sequence into protein. This is done by:

right-click the tab of the new sequence | **Toolbox** | **Nucleotide Analyses (** **)** | **Translate to Protein (** **)**

A selection can also be copied to the clipboard and pasted into another program:

make a selection | **Ctrl + C (⌘ + C on Mac)**

Notice! The annotations covering the selection will not be copied.

A selection of a sequence can be edited as described in the following section.

11.1.3 Editing the sequence

When you make a selection, it can be edited by:

right-click the selection | **Edit selection**

A dialog appears displaying the sequence. You can add, remove or change the text and click **OK**. The original selected part of the sequence is now replaced by the sequence entered in the dialog. This dialog also allows you to paste text into the sequence using Ctrl + V (⌘ + V on Mac). If you delete the text in the dialog and press **OK**, the selected text on the sequence will also be deleted. Another way to delete a part of the sequence is to:

right-click the selection | **Delete selection**

Another way to edit the sequence is by inserting a restriction site. See section 17.2.2.

11.1.4 Adding and modifying annotations

Most sequences carry different biological information. When retrieving sequences from various databases, the sequence often contains biological information by way of annotations. You can manually add annotations from a compiled annotation list. This list of annotations covers the most frequently used annotations in UniProt and GenBank. Annotations which have been added to a sequence can be removed at any time (see section 11.1.5).

Annotations can be added to a sequence:

make a selection covering the part of the sequence you want to annotate | **right-click the selection** | **Add Annotation**

This will display a dialog like the one in figure 11.2.

The left-hand part of the dialog lists a number of **Annotation types**. When you have selected an annotation type, it appears in **Chosen type**. You can also select an annotation from the **Chosen type** list. Choosing an annotation type is mandatory.

The right-hand part of the dialog contains the following text fields:

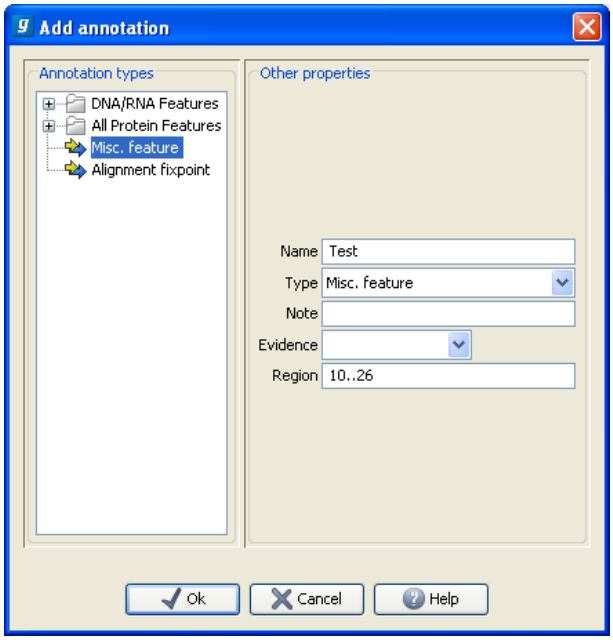


Figure 11.2: *The Add Annotation dialog.*

- **Name.** The name of the annotation which can be shown in the view. Whether the name is shown depends on the **Annotation Layout** preferences (see section 11.1.1).
- **Chosen type.** Reflects the left-hand part of the dialog as described above.
- **Note.** This is a field for entering notes about the annotation. The note will be displayed in a tooltip when you hold the mouse pointer over the sequence.
- **Evidence.** There are two options for the evidence supporting the annotation: experimental and non-experimental.
- **Region.** If you have already made a selection, this field will show the positions of the selection. You can modify the region further using the syntax of using the conventions of DDBJ, EMBL and GenBank. The following are examples of how to use the syntax (based on `http://www.ncbi.nlm.nih.gov/collab/FT/`):
  - **467**. Points to a single residue in the presented sequence.
  - **340..565**. Points to a continuous range of residues bounded by and including the starting and ending residues.
  - **<345..500**. Indicates that the exact lower boundary point of a region is unknown. The location begins at some residue previous to the first residue specified (which is not necessarily contained in the presented sequence) and continues up to and including the ending residue.
  - **<1..888**. The region starts before the first sequenced residue and continues up to and including residue 888.
  - **1..>888**. The region starts at the first sequenced residue and continues beyond residue 888.
  - **(102.110)**. Indicates that the exact location is unknown, but that it is one of the residues between residues 102 and 110, inclusive.

- **123^124**. Points to a site between residues 123 and 124.
- **join(12..78,134..202)**. Regions 12 to 78 and 134 to 202 should be joined to form one contiguous sequence.
- **complement(34..126)** Start at the residue complementary to 126 and finish at the residue complementary to residue 34 (the region is on the strand complementary to the presented strand).
- **complement(join(2691..4571,4918..5163))**. Joins regions 2691 to 4571 and 4918 to 5163, then complements the joined segments (the region is on the strand complementary to the presented strand).
- **join(complement(4918..5163),complement(2691..4571))**. Complements regions 4918 to 5163 and 2691 to 4571, then joins the complemented segments (the region is on the strand complementary to the presented strand).

Click **OK** to add the annotation.

Notice! The annotation will be included if you export the sequence in GenBank, Swiss-Prot or CLC format.

To modify an existing annotation:

right-click the annotation | **Edit Annotation**

This will show the same dialog as in figure 11.2, with the exception that some of the fields are filled out depending on how much information the annotation contains.

11.1.5 Removing annotations

Annotations can be hidden using the **Annotation Types** preferences in the **Side Panel** to the right of the view (see section 11.1.1). In order to completely remove the annotation:

right-click the annotation | **Delete Annotation**

If you want to remove all annotations of one type:

right-click an annotation of the type you want to remove | **Delete Annotations of This Type**

If you want to remove all annotations from a sequence:

right-click an annotation | **Delete All Annotations**

The removal of annotations can be undone using Ctrl + Z or Undo ( ) in the Toolbar.

11.1.6 Sequence region types

The various annotations on sequences cover parts of the sequence. Some cover an interval, some cover intervals with unknown endpoints, some cover more than one interval etc. In the following, all of these will be referred to as *regions*. Regions are generally illustrated by markings (often arrows) on the sequences. An arrow pointing to the right indicates that the corresponding region is located on the positive strand of the sequence. Figure 11.3 is an example of three regions with separate colors.

Figure 11.4 shows an artificial sequence with all the different kinds of regions.

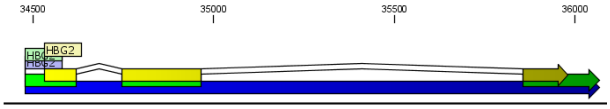


Figure 11.3: *Three regions on a human beta globin DNA sequence (HUMHBB).*

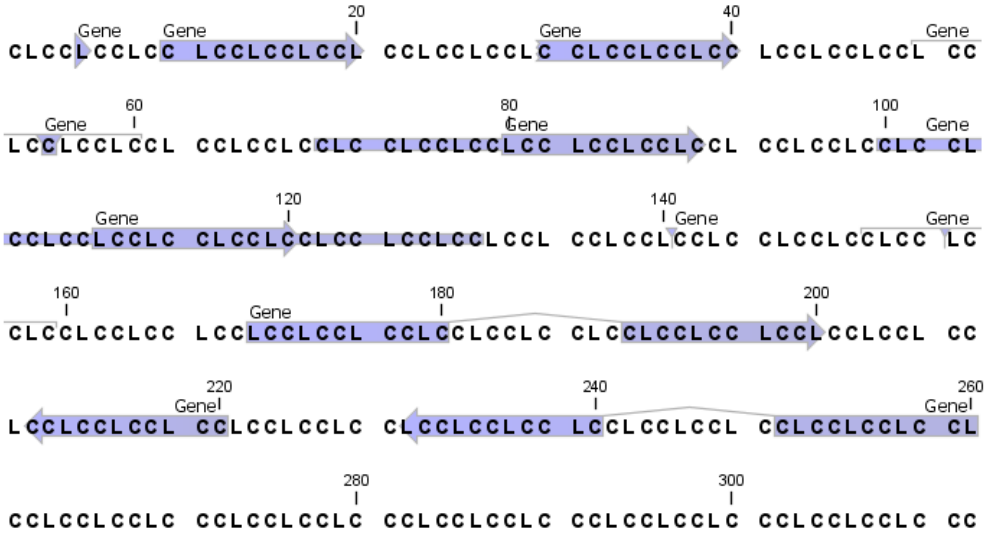


Figure 11.4: *Region #1: A single residue, Region #2: A range of residues including both endpoints, Region #3: A range of residues starting somewhere before 30 and continuing up to and including 40, Region #4: A single residue somewhere between 50 and 60 inclusive, Region #5: A range of residues beginning somewhere between 70 and 80 inclusive and ending at 90 inclusive, Region #6: A range of residues beginning somewhere between 100 and 110 inclusive and ending somewhere between 120 and 130 inclusive, Region #7: A site between residues 140 and 141, Region #8: A site between two residues somewhere between 150 and 160 inclusive, Region #9: A region that covers ranges from 170 to 180 inclusive and 190 to 200 inclusive, Region #10: A region on negative strand that covers ranges from 210 to 220 inclusive, Region #11: A region on negative strand that covers ranges from 230 to 240 inclusive and 250 to 260 inclusive.*

11.2 Sequence information

The normal view of a sequence (by double-clicking) shows the annotations as boxes along the sequence, but often there is more information available about sequences. This information is available through the **Sequence info** function which also displays a textual overview of the annotations.

To view the sequence information:

> **select a sequence in the Navigation Area | Show ( ) in the Toolbar | Sequence info ( )**

This will display a view similar to fig 11.5.

All the lines in the view are headings, and the corresponding text can be shown by clicking the text. The information available depends on the origin of the sequence. If the sequence is annotated, the annotations can be found under the heading **Annotation map**.

Figure 11.5: *The initial display of sequence info for the HUMHBB DNA sequence from the Example data.*

11.2.1 Annotation map

The **Annotation map** displays the various types of annotations that are attached to the sequence. Clicking on the name of a type of annotation will list the annotations of this type. If there are more annotations of the same kind, the blue arrows can be used to move up and down in the annotations of that type. In order to use the links, you have to open a second view of the sequence (double-click the sequence in the **Navigation Area**). If you have this view open, clicking one of the annotations in the **Annotation map** will make a selection in the other view corresponding to the annotation (see fig 11.6).

Annotations cannot be added or modified using the **Sequence info**. For adding and modifying annotations see section 11.1.4.

11.3 View as text

A sequence can be viewed as text without any layout and text formatting. This displays all the information about the sequence in the GenBank file format. To view a sequence as text:

select a sequence in the Navigation Area | Show in the Toolbar | As text

This way it is possible to see background information about e.g. the authors and the origin of DNA and protein sequences. Selections or the entire text of the **Sequence Text Viewer** can be copied and pasted into other programs:

Much of the information is also displayed in the **Sequence info**, where it is easier to get an overview (see section 11.2.)

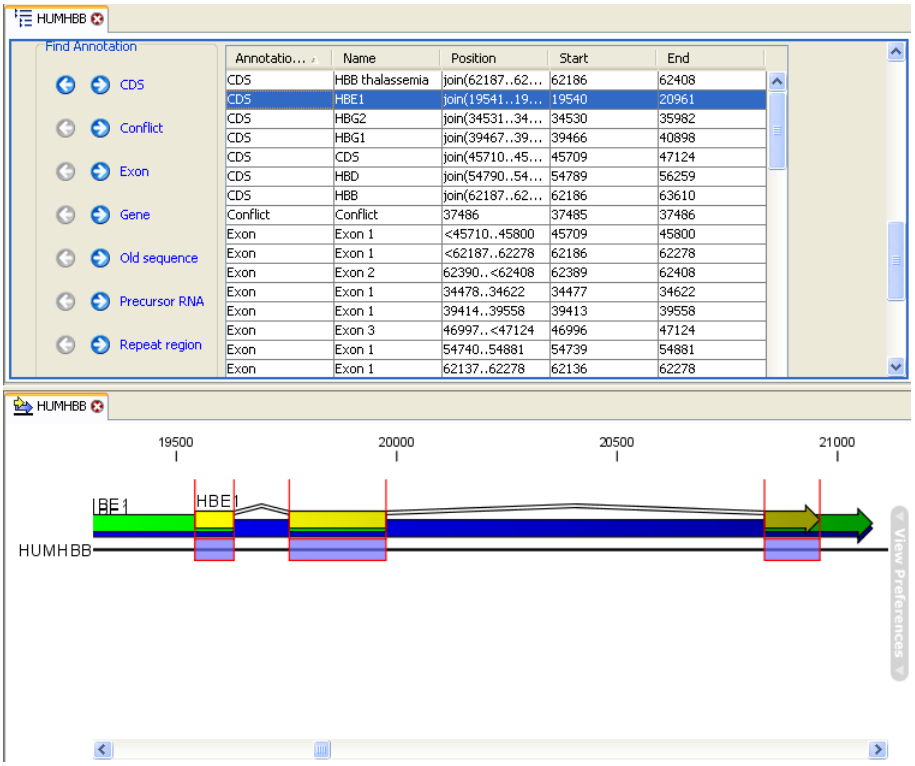


Figure 11.6: *Clicking a sequence map annotation in the sequence information view, selects the annotation on the (normal) sequence view.*

11.4 Creating a new sequence

A sequence can either be imported, downloaded from an online database or created in the *CLC Gene Workbench 2.0*. This section explains how to create a new sequence:

New( ) in the toolbar

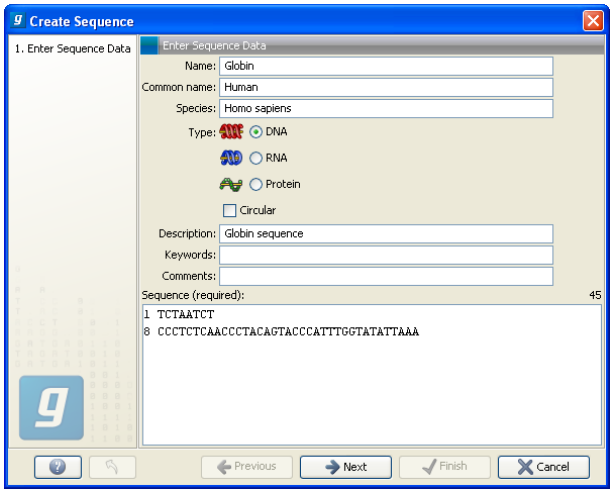


Figure 11.7: *Creating a sequence.*

The **Create Sequence** dialog (figure 11.7) reflects the information needed in the GenBank format, but you are free to enter anything into the fields. The following description is a guideline for entering information about a sequence:

- **Name.** The name of the sequence. This is used for saving the sequence.
- **Common name.** A common name for the species.
- **Species.** The Latin name.
- **Type.** Select between DNA, RNA and protein.
- **Circular.** Specifies whether the sequence is circular. This will open the sequence in a circular view as default. (applies only to nucleotide sequences).
- **Description.** A description of the sequence.
- **Keywords.** A set of keywords separated by semicolons (;).
- **Comments.** Your own comments to the sequence.
- **Sequence.** Depending on the type chosen, this field accepts nucleotides or amino acids. Spaces and numbers can be entered, but they are ignored when the sequence is created. This allows you to paste in a sequence directly from a different source, even if the residue numbers are included. Characters that are not part of the IUPAC codes cannot be entered. At the top right corner of the field, the number of residues are counted. The counter does not count spaces or numbers.

Clicking Next will allow you to save the sequence to a project in the **Navigation Area**.

11.5 Sequence Lists

The **Sequence List** shows a number of sequences in a tabular format or it can show the sequences together in a normal sequence view.

Having sequences in a sequence list can help organizing sequence data. The sequence list may originate from an NCBI search (chapter 9.1). Moreover, if a multiple sequence fasta file is imported, it is possible to store the data in a sequences list. A **Sequence List** can also be generated using a dialog, which is described here:

select two or more sequences | right-click the elements | New | Sequence List ( )

This action opens a **Sequence List** dialog:

The dialog allows you to select more sequences to include in the list, or to remove already chosen sequences from the list.

After clicking "Next", you can choose where to save the list. Then click **Finish**.

Opening a Sequence list is done by:

right-click the sequence list in the Navigation Area | Show | click Graphical sequence list OR click Table

The two different views of the same sequence list are shown in split screen in figure 11.9.

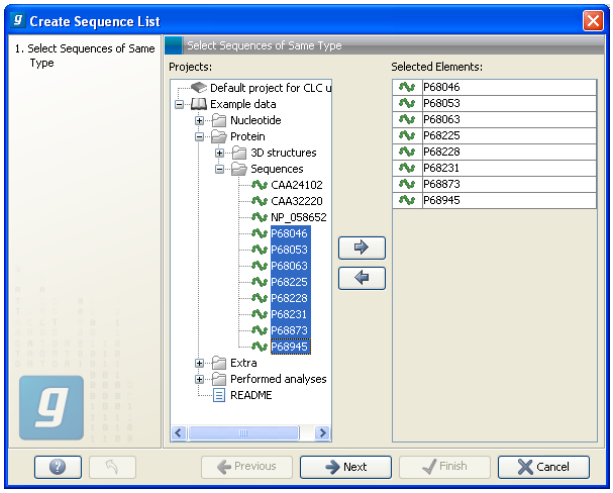


Figure 11.8: *A Sequence List dialog.*

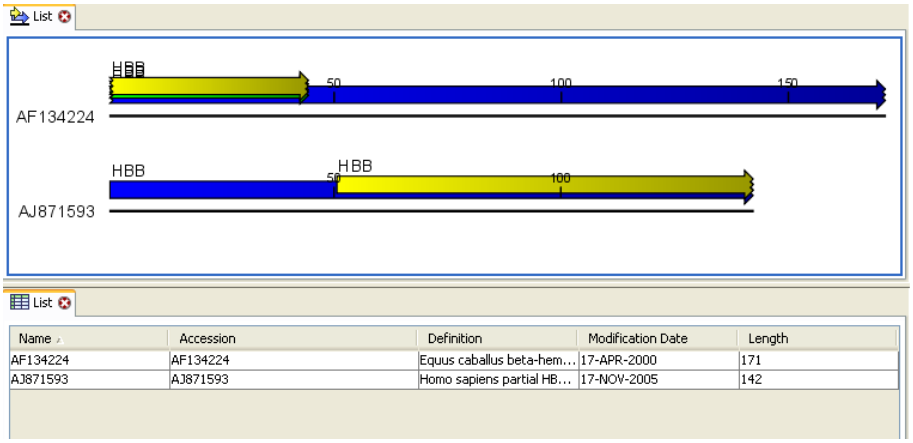


Figure 11.9: *A sequence list of two sequences can be viewed in either a table or in a graphical sequence list.*

11.5.1 Graphical view of sequence lists

The graphical view of sequence lists is almost identical to the view of single sequences (see section 11.1). The main difference is that you now can see more than one sequence in the same view.

However, you also have a few extra options for sorting, deleting and adding sequences:

- To add extra sequences to the list, right-click an empty (white) space in the view, and select **Add Sequences**.
- To delete a sequence from the list, right-click the sequence's label and select **Delete Sequence**.
- To sort the sequences in the list, right-click the label of one of the sequences and select **Sort Sequence List by Name** or **Sort Sequence List by Length**.
- To rename a sequence, right-click the label of the sequence and select **Rename Sequence**.

11.5.2 Sequence list table

Each sequence in the table sequence list is displayed with:

- Name.
- Accession.
- Definition.
- Modification date.
- Length.

In the View preferences for the table view of the sequence list, columns can be excluded, and the view preferences can be saved in a style sheet. See section 4.5.

The sequences can be sorted by clicking the column headings. You can further refine the sorting by pressing Ctrl while clicking the heading of another column.

11.5.3 Extract sequences

It is possible to extract individual sequences from a sequence list in two ways. If the sequence list is opened in the tabular view, it is possible to drag (with the mouse) one or more sequences into the **Navigation Area**. This allows you to extract specific sequences from the entire list. Another option is to extract all sequences found in the list to a preferred location in the **Navigation Area**:

right-click a sequence list in the Navigation Area | Extract Sequences

Select a location for the sequences and click OK. Copies of all the sequences in the list are now placed in the location you selected.

11.6 Circular DNA

A sequence can be shown as a circular molecule:

select a sequence in the Navigation Area | Show in the Toolbar | Circular( )

This will open a view of the molecule similar to the one in figure 11.10.

This view of the sequence shares some of the properties of the linear view of sequences as described in section 11.1, but there are some differences. The similarities and differences are listed below:

- **Similarities**:
  - The editing options.
  - Options for adding, editing and removing annotations.
  - **Annotation Layout**, **Annotation Types** and **Text Format** preferences groups.
- **Differences**:

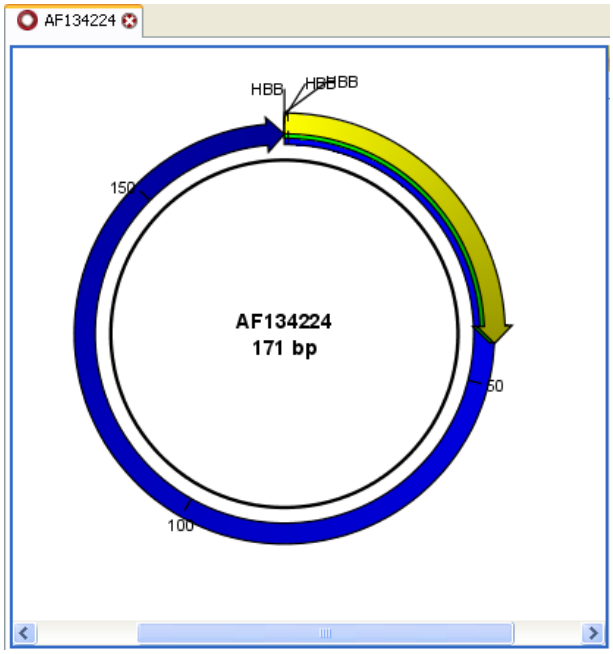


Figure 11.10: *A molecule shown in a circular view.*

- In the **Sequence Layout** preferences, only the following options are available in the circular view: **Numbers on plus strand**, **Numbers on sequence** and **Sequence label**.
- The circular view can display restriction sites using the **Restriction Sites** preference group (described below).
- You cannot zoom in to see the residues in the circular molecule. If you wish to see these details, split the view with a linear view of the sequence (described below).

11.6.1 Show restriction sites for circular DNA

These preferences allow you to display restriction sites on the sequence. There is a list of enzymes which are represented by different colors. By selecting or deselecting the enzymes in the list, you can specify which enzymes' restriction sites should be displayed (see figure 17.4).

Figure 11.11: *Showing restriction sites of two restriction enzymes.*

The color of the flag of the restriction site can be changed by clicking the colored box next to the enzyme's name.

The list of restriction enzymes contains per default ten of the most popular enzymes, but you can easily modify this list and add more enzymes. You have four ways of modifying the list:

- **Edit enzymes button.** This displays a dialog with the enzymes currently in the list shown at the bottom and a list of available enzymes at the top. To add more enzymes, select them in the upper list and press the **Add enzymes button** (). To remove enzymes, select them in the list below and click the **Remove enzymes button** ().

- **Load enzymes button.** If you have previously created an enzyme list, you can select this list by clicking the Load enzymes button. You can filter the enzymes in the same way as illustrated in figure 17.13.
- **Add enzymes cutting the selection to panel.** If you make a selection on the sequence, right-click, you find this option for adding enzymes. Based on the entire list of available enzymes, the enzymes cutting in the region you selected will be added to the list in the **Side Panel**.
- **Insert restriction site before/after selection.** If you make a selection on the sequence, right-click, you find this option for inserting a restriction site before or after the region you selected. A dialog is shown where you can select an enzyme whose recognition sequence is inserted. If it was not already present in the list in the **Side Panel**, the enzyme will now be added and selected.

Finally, if you have selected a set of enzymes that you wish to keep for later use, you can click **Save enzymes** and the selected enzymes will be saved to en enzyme list. This list can then be used both when finding restriction sites from the **Toolbox** or when viewing another sequence.

11.6.2 Using split views to see details of the circular molecule

In order to see the nucleotides of a circular molecule you can open a new view displaying a circular view of the molecule:

right-click the tab of the circular view of the sequence | Show | Sequence( )

This will open a linear view of the sequence below the circular view. When you zoom in on the linear view you can see the residues as shown in figure 11.12.

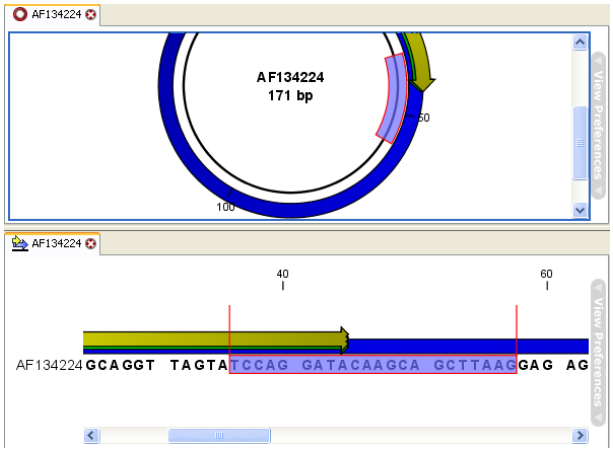


Figure 11.12: *Two views showing the same sequence. The bottom view is zoomed in.*

Notice! If you make a selection in one of the views, the other view will also make the corresponding selection, providing an easy way for you to focus on the same region in both views.

11.6.3 Mark molecule as circular and specify starting point

You can mark a DNA molecule as circular by right-clicking its label in either the sequence view or the circular view. In the right-click menu you can also make a circular molecule linear. A circular molecule displayed in the normal sequence view, will have the sequence ends marked with a ».

The starting point of a circular sequence can be changed by:

> **make a selection starting at the position that you want to be the new starting point | right-click the selection | Move Starting Point to Selection Start**

Notice! This can only be done for sequence that have been marked as circular.

Chapter 12

General sequence analyses

Contents

12.1 Dot plots . . . 136
12.1.1 Create dot plots . . . 137
12.1.2 View dot plots . . . 138
12.1.3 Bioinformatics explained: Dot plots . . . 139
12.1.4 Bioinformatics explained: Scoring matrices . . . 142
12.2 Shuffle sequence . . . 146
12.3 Local complexity plot . . . 147
12.3.1 Local complexity view preferences . . . 148
12.4 Sequence statistics . . . 149
12.4.1 Sequence statistics output . . . 152
12.4.2 Bioinformatics explained: Protein statistics . . . 152
12.5 Join sequences . . . 156
12.6 Motif Search . . . 157
12.6.1 Motif search parameter settings . . . 159
12.6.2 Motif search output . . . 160
12.7 Pattern Discovery . . . 160
12.7.1 Pattern discovery search parameters . . . 161
12.7.2 Pattern search output . . . 162

CLC Gene Workbench 2.0 offers different kinds of sequence analyses, which apply to both protein and DNA. The analyses are described in this chapter.

12.1 Dot plots

Dot plots provide a powerful visual comparison of two sequences. Dot plots can also be used to compare regions of similarity within a sequence. This chapter first describes how to create and second how to adjust the view of the plot.

12.1.1 Create dot plots

A dot plot is a simple, yet intuitive way of comparing two sequences, either DNA or protein, and is probably the oldest way of comparing two sequences [Maizel and Lenk, 1981]. A dot plot is a 2 dimensional matrix where each axis of the plot represents one sequence. By sliding a fixed size window over the sequences and making a sequence match by a dot in the matrix, a diagonal line will emerge if two identical (or very homologous) sequences are plotted against each other. Dot plots can also be used to visually inspect sequences for direct or inverted repeats or regions with low sequence complexity. Various smoothing algorithms can be applied to the dot plot calculation to avoid noisy background of the plot. Moreover, can various substitution matrices be applied in order to take the evolutionary distance of the two sequences into account.

To create a dot plot:

Toolbox | **General Sequence Analyses ( )** | **Create Dot Plot( )**

or **Select one or two sequences in the Navigation Area** | **Toolbox in the Menu Bar** | **General Sequence Analyses ( )** | **Create Dot Plot ( )**

or **Select one or two sequences in the Navigation Area** | **right-click in the Navigation Area** | **Toolbox** | **General Sequence Analyses ( )** | **Create Dot Plot( )**

This opens the dialog shown in figure 12.1.

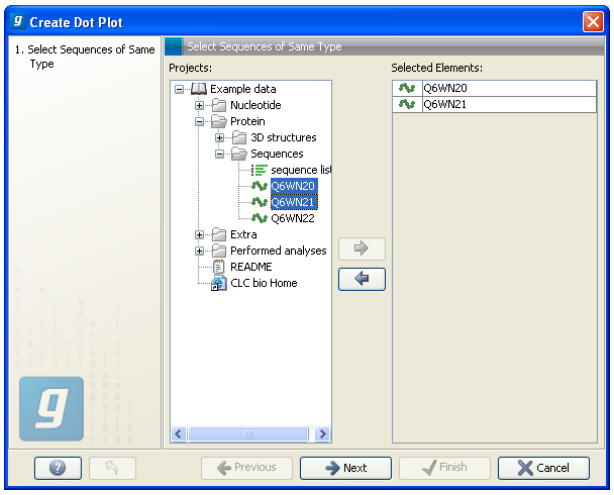


Figure 12.1: *Selecting sequences for the dot plot.*

If a sequence was selected before choosing the **Toolbox** action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove elements from the **Project Tree**. Click **Next** to adjust dot plot parameters. Clicking **Next** opens the dialog shown in figure 12.2.

Notice that calculating dot plots take up a considerable amount of memory in the computer. Therefore, you see a warning if the sum of the number of nucleotides/amino acids in the sequences is higher than 8000. If you insist on calculating a dot plot with more residues the Workbench may shut down, allowing you to save your work first. To avoid the Workbench shutting down you may choose to adjust the memory allocation to *CLC Gene Workbench*.
See section 1.8

Adjust dot plot parameters

There are two parameters for calculating the dot plot:

- **Distance correction (only valid for protein sequences)** In order to treat evolutionary transitions of amino acids, a distance correction measure can be used when calculating the dot plot. These distance correction matrices (substitution matrices) take into account the likeliness of one amino acid changing to another.
- **Window size** A residue by residue comparison (window size = 1) would undoubtedly result in a very noisy background due to a lot of similarities between the two sequences of interest. For DNA sequences the background noise will be even more dominant as a match between only four nucleotide is very likely to happen. Moreover, a residue by residue comparison (window size = 1) can be very time consuming and computationally demanding. Increasing the window size will make the dot plot more 'smooth'.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

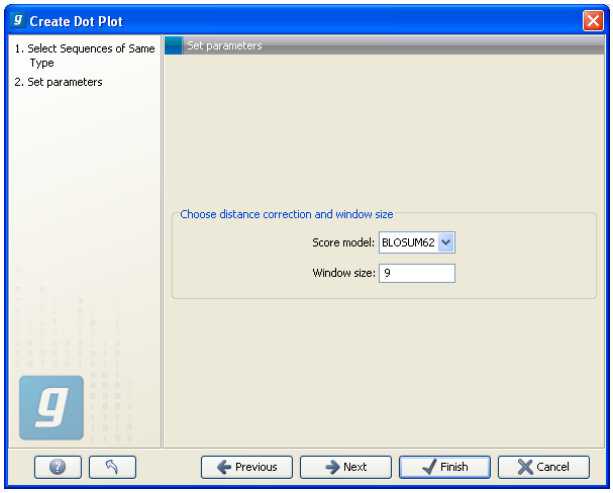


Figure 12.2: *Setting the dot plot parameters.*

12.1.2 View dot plots

A view of a dot plot can be seen in figure 12.3. You can select **Zoom in** ( ) in the Toolbar and click the dot plot to zoom in to see the details of particular areas.

The **Side Panel** to the right let you specify the dot plot preferences. The gradient color box can be adjusted to get the appropriate result by dragging the small pointers at the top of the box. Moving the slider from the right to the left lowers the thresholds which can be directly seen in the dot plot, where more diagonal lines will emerge. You can also choose another color gradient by clicking on the gradient box and choose from the list.

Adjusting the sliders above the gradient box is also practical, when producing an output for printing. (Too much background color might not be desirable). By crossing one slider over the other (the two sliders change side) the colors are inverted, allowing for a white background. (If you choose a color gradient, which includes white). Se figure 12.3.

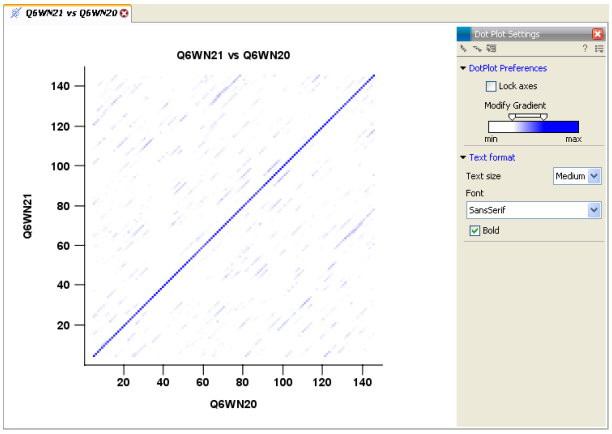


Figure 12.3: *A view is opened showing the dot plot.*

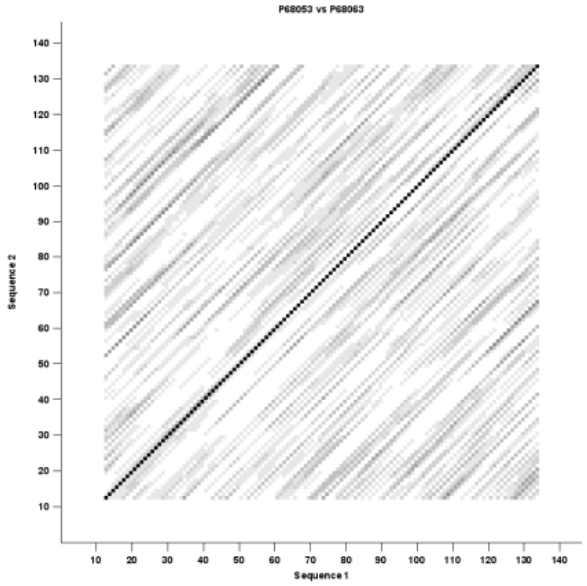


Figure 12.4: *Dot plot with inverted colors, practical for printing.*

12.1.3 Bioinformatics explained: Dot plots

Realization of dot plots

Dot plots are two-dimensional plots where the x-axis and y-axis each represents a sequence and the plot itself shows a comparison of these two sequences by a calculated score for each position of the sequence. If a window of fixed size on one sequence (one axis) match to the other sequence a dot is drawn at the plot. Dot plots are one of the oldest methods for comparing two sequences [Maizel and Lenk, 1981].

The scores that are drawn on the plot are affected by several issues.

- Scoring matrix for distance correction.
  Scoring matrices (BLOSUM and PAM) contain substitution scores for every combination of two amino acids. Thus, these matrices can only be used for dot plots of protein sequences.
- Window size
  The single residue comparison (bit by bit comparison(window size = 1)) in dot plots will

undoubtedly result in a noisy background of the plot. You can imagine that there are many successes in the comparison if you only have four possible residues like in nucleotide sequences. Therefore you can set a window size which is smoothing the dot plot. Instead of comparing single residues it compares subsequences of length set as window size. The score is now calculated with respect to aligning the subsequences.

- Threshold
  The dot plot shows the calculated scores with colored threshold. Hence you can better recognize the most important similarities.

Examples and interpretations of dot plots

Contrary to simple sequence alignments dot plots can be a very useful tool for spotting various evolutionary events which may have happened to the sequences of interest.

Below is shown some examples of dot plots where sequence insertions, low complexity regions, inverted repeats etc. can be identified visually.

Similar sequences

The most simple example of a dot plot is obtained by plotting two homologous sequences of interest. If very similar or identical sequences are plotted against each other a diagonal line will occur.

The dot plot in figure 12.5 shows two related sequences of the Influenza A virus nucleoproteins infecting ducks and chickens. Accession numbers from the two sequences are: DQ232610 and DQ023146. Both sequences can be retrieved directly from `http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi`.

Repeated regions

Sequence repeats can also be identified using dot plots. A repeat region will typically show up as lines parallel to the diagonal line.

If the dot plot shows more than one diagonal in the same region of a sequence, the regions depending to the other sequence are repeated. In figure 12.7 you can see a sequence with repeats.

Frame shifts

Frame shifts in a nucleotide sequence can occur due to insertions, deletions or mutations. Such frame shifts can be visualized in a dot plot as seen in figure 12.8. In this figure, three frame shifts for the sequence on the y-axis are found.

1. Deletion of nucleotides
2. Insertion of nucleotides
3. Mutation (out of frame)

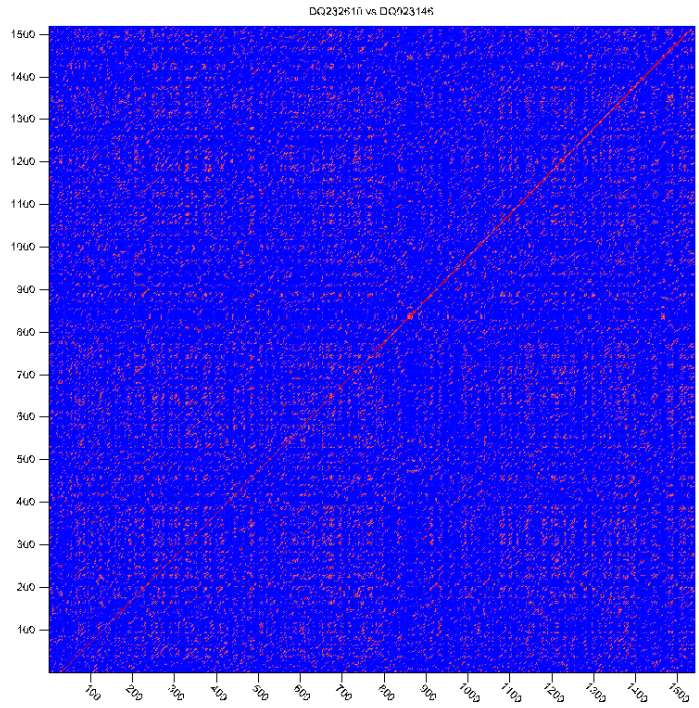


Figure 12.5: *Dot plot of DQ232610 vs. DQ023146 (Influenza A virus nucleoproteins) showing and overall similarity*

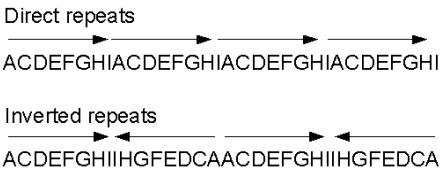


Figure 12.6: *Direct and inverted repeats shown on an amino acid sequence generated for demonstration purposes.*

Sequence inversions

In dot plots you can see an inversion of sequence as contrary diagonal to the diagonal showing similarity. In figure 12.9 you can see a dot plot (window length is 3) with an inversion.

Low-complexity regions

Low-complexity regions in sequences can be found as regions around the diagonal all obtaining a high score. Low complexity regions are calculated from the redundancy of amino acids within a limited region [Wootton and Federhen, 1993]. These are most often seen as short regions of only a few different amino acids. In the middle of figure 12.10 is a square shows the low-complexity region of this sequence.

Creative Commons License
All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and

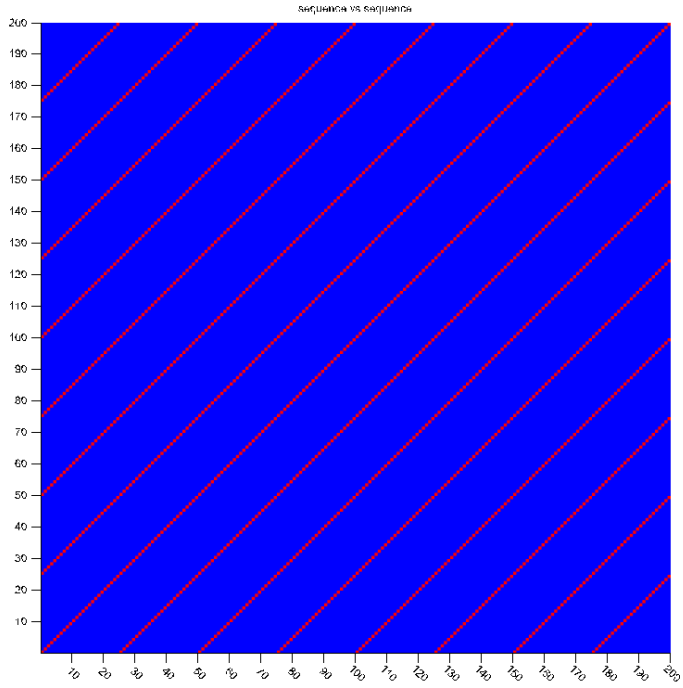


Figure 12.7: *The dot plot of a sequence showing repeated elements. See also figure 12.6.*

"CLC bio" has to be clearly labelled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.

SOME RIGHTS RESERVED

See http://creativecommons.org/licenses/by-nc-nd/2.5/ for more about how you may use the contents.

12.1.4 Bioinformatics explained: Scoring matrices

Biological sequences have evolved throughout time and evolution has shown that not all changes to a biological sequence is equally likely to happen. Certain amino acid substitutions (change of one amino acid to another) happen often, whereas other substitutions are very rare. For instance, tryptophan (W) which is a relatively rare amino acid, will only — on very rare occasions — mutate into a leucine (L).

Based on evolution of proteins it became apparent that these changes or substitutions of amino acids can be modeled by a scoring matrix also refereed to as a substitution matrix. See an example of a scoring matrix in table 12.1. This matrix lists the substitution scores of every single amino acid. A score for an aligned amino acid pair is found at the intersection of the corresponding column and row. For example, the substitution score from an arginine (R) to

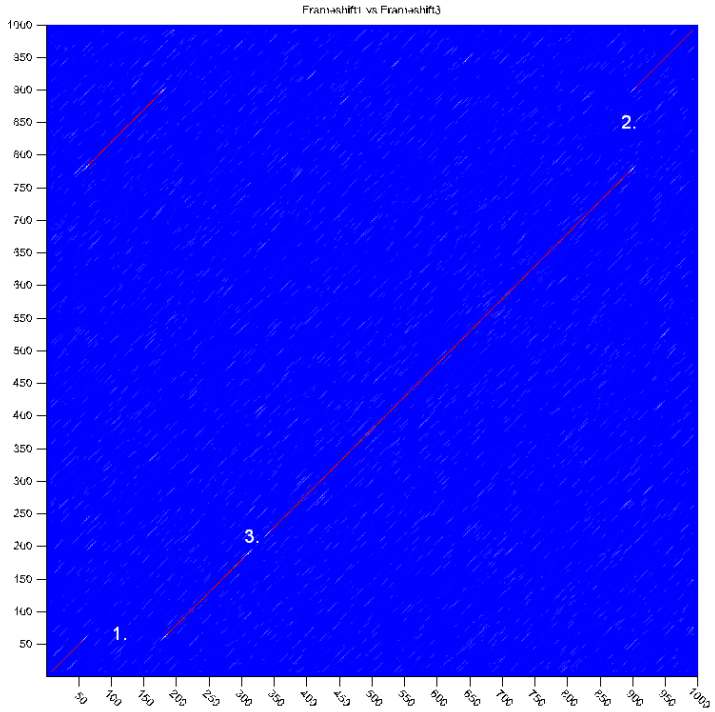


Figure 12.8: *This dot plot show various frame shifts in the sequence. See text for details.*

a lysine (K) is 2. The diagonal show scores for amino acids which have not changed. Most substitutions changes have a negative score. Only rounded numbers are found in this matrix.

The two most used matrices are the BLOSUM [Henikoff and Henikoff, 1992] and PAM [Dayhoff and Schwartz, 1978].

Different scoring matrices

PAM

The first PAM matrix (Point Accepted Mutation) was published in 1978 by Dayhoff et al. The PAM matrix was build through a global alignment of related sequences all having sequence similarity above 85% [Dayhoff and Schwartz, 1978]. A PAM matrix shows the probability that any given amino acid will mutate into another in a given time interval. As an example, PAM1 gives that one amino acid out of a 100 will mutate in a given time interval. In the other end of the scale, a PAM256 matrix, gives the probability of 256 mutations in a 100 amino acids (see figure 12.11).

There are some limitation to the PAM matrices which makes the BLOSUM matrices somewhat more attractive. The dataset on which the initial PAM matrices were build is very old by now, and the PAM matrices assume that all amino acids mutate at the same rate - this is not a correct assumption.

BLOSUM

In 1992, 14 years after the PAM matrices were published, the BLOSUM matrices (BLOcks SUbstitution Matrix) were developed and published [Henikoff and Henikoff, 1992].

Henikoff et al. wanted to model more divergent proteins, thus they used locally aligned

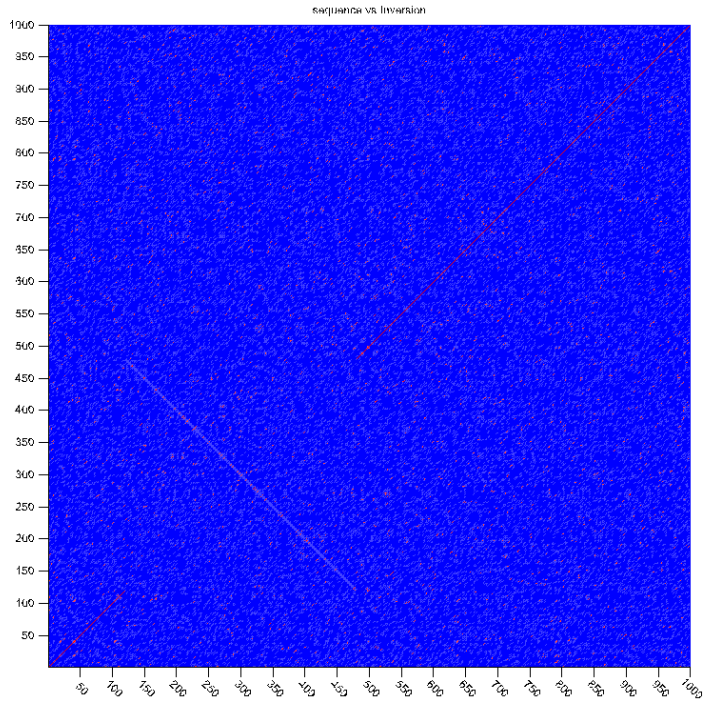


Figure 12.9: *The dot plot showing a inversion in a sequence. See also figure 12.6.*

sequences where none of the aligned sequences share less than 62% identity. This resulted in a scoring matrix´called BLOSUM62. In contrast to the PAM matrices the BLOSUM matrices are calculated from alignments without gaps emerging from the BLOCKS database `http://blocks.fhcrc.org/`.

Sean Eddy recently wrote a paper reviewing the BLOSUM62 substitution matrix and how to calculate the scores [Eddy, 2004].

Use of scoring matrices

Deciding which scoring matrix you should use in order of obtain the best alignment results is a difficult task. If you have no prior knowledge on the sequence the BLOSUM62 is probably the best choice. This matrix has become the *de facto* standard for scoring matrices and is also used as the default matrix in BLAST searches. The selection of a "wrong" scoring matrix will most probable strongly influence on the outcome of the analysis. In general a few rules apply to the selection of scoring matrices.

- For closely related sequences choose BLOSUM matrices created for highly similar alignments, like BLOSUM80. You can also select low PAM matrices such as PAM1.
- For distant related sequences, select low BLOSUM matrices (for example BLOSUM45) or high PAM matrices such as PAM250.

The BLOSUM matrices with low numbers correspond to PAM matrices with high numbers. (See figure 12.11) for correlations between the PAM and BLOSUM matrices. To summarize, if you

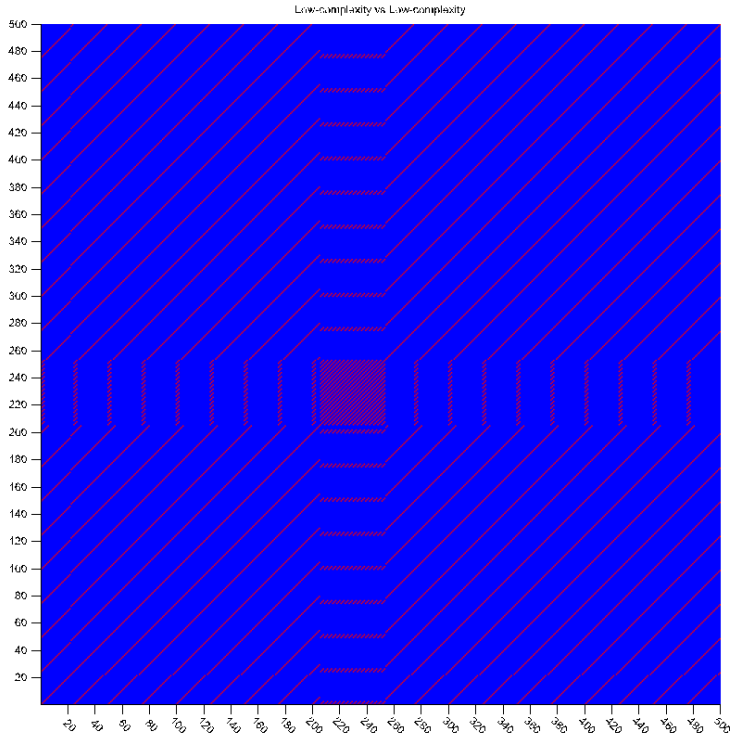


Figure 12.10: *The dot plot showing a low-complexity region in the sequence. The sequence is artificial and low complexity regions does not always show as a square.*

want to find distant related proteins to a sequence of interest using BLAST, you could benefit of using BLOSUM45 or similar matrices.

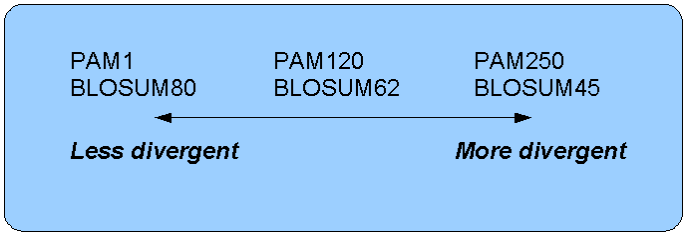


Figure 12.11: *Relationship between scoring matrices. The BLOSUM62 has become a* de facto *standard scoring matrix for a wide range of alignment programs. It is the default matrix in BLAST.*

Other useful resources

Calculate your own PAM matrix
`http://www.bioinformatics.nl/tools/pam.html`

BLOKS database
`http://blocks.fhcrc.org/`

NCBI help site
`http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Scoring2.html`

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Table 12.1: **The BLOSUM62 matrix**. A tabular view of the BLOSUM62 matrix containing all possible substitution scores [Henikoff and Henikoff, 1992].

Creative Commons License
All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and "CLC bio" has to be clearly labelled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.

See http://creativecommons.org/licenses/by-nc-nd/2.5/ for more about how you may use the contents.

12.2 Shuffle sequence

In some cases, it is beneficial to shuffle a sequence. This is an option in the **Toolbox** menu under **General Sequence Analyses**. It is normally used for statistical analyses, e.g. when comparing an alignment score with the distribution of scores of shuffled sequences. The shuffling is done without replacement, resulting in exactly the same number of the different residues as before the shuffling.

Shuffling a sequence removes all annotations that relate to the residues.

select sequence | **Toolbox in the Menu Bar** | **General Sequence Analyses** ( ) | **Shuffle Sequence** ( )

or **right-click a sequence** | **Toolbox** | **General Sequence Analyses** ( ) | **Shuffle Sequence** ( )

This opens the dialog displayed in figure 12.12:

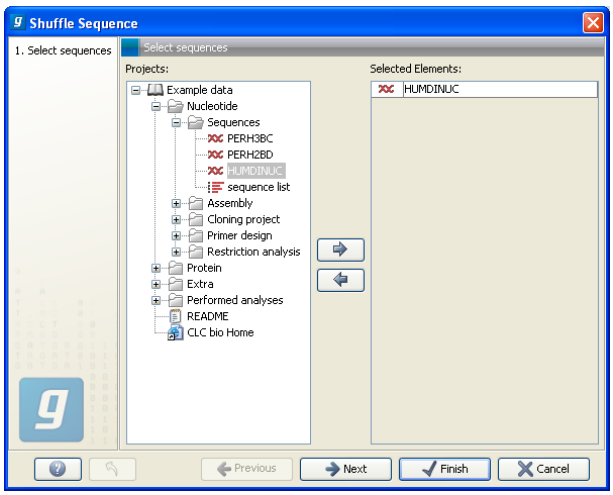


Figure 12.12: *Choosing sequence for shuffling.*

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the **Project Tree**.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

This will open a new view in the **View Area** displaying the shuffled sequence. The new sequence is not saved automatically. To save the protein sequence, drag it into the **Navigation Area** or press ctrl + S (⌘ + S on Mac) to activate a save dialog.

12.3 Local complexity plot

In *CLC Gene Workbench* it is possible to calculate local complexity for both DNA and protein sequences. The local complexity is a measure of the diversity in the composition of amino acids within a given range (window) of the sequence. The K2 algorithm is used for calculating local complexity [Wootton and Federhen, 1993]. To conduct a complexity calculation do the following:

Select sequences in Navigation Area | **Toolbox in Menu Bar** | **General Sequence Analyses** ( ) | **Create Complexity Plot** ( )

This opens a dialog. In **Step 1** you can change, remove and add DNA and protein sequences.

When the relevant sequences are selected, clicking **Next** takes you to **Step 2**. This step allows you to adjust the window size from which the complexity plot is calculated. Default is set to 11 amino acids and the number should always be odd. The higher the number, the less volatile the graph.

Figure 12.13 shows an example of af local complexity plot.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. The values of the complexity plot approaches 1.0 as the distribution of amino acids become more complex.

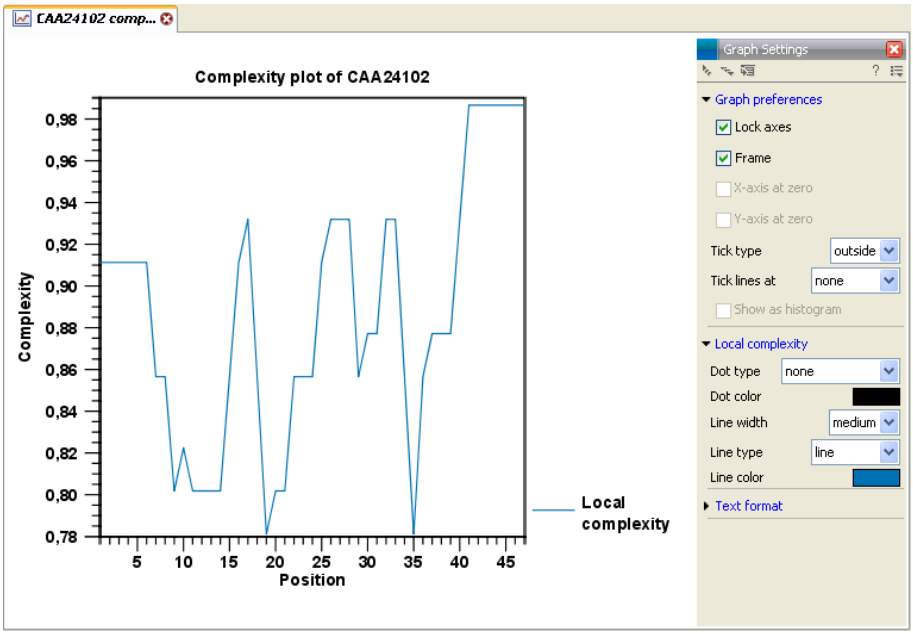


Figure 12.13: *An example of a local complexity plot.*

12.3.1 Local complexity view preferences

There are two groups of preferences for the local complexity view: Graph preferences and Local complexity preferences:

The **Graph preferences** apply to the whole graph:

- **Lock axis.**
  This will always show the axis even though the plot is zoomed to a detailed level.
- **Frame.**
  Toggles the frame of the graph.
- **X-axis at zero.**
  Toggles the x-axis at zero.
- **Y-axis at zero.**
  Toggles the y-axis at zero.
- **Tick type**
  - outside
  - inside
- **Tick lines at.**
  Shows a grid behind the graph.
  - none
  - major ticks
- **Show as histogram**
  For some data-series it is possible to see it as a histogram rather than a line plot.

The **Local complexity preferences** include:

- **Dot type**
  - none
  - cross
  - plus
  - square
  - diamond
  - circle
  - triangle
  - reverse triangle
  - dot
- **Dot color.** Allows you to choose between many different colors.
- **Line width**
  - thin
  - medium
  - wide
- **Line type**
  - none
  - line
  - long dash
  - short dash
- **Line color.** Allows you to choose between many different colors.

12.4 Sequence statistics

CLC Gene Workbench 2.0 can produce an output with many relevant statistics for protein sequences. Some of the statistics are also relevant to produce for DNA sequences. Therefore, this section deals with both types of statistics. The required steps for producing the statistics are the same.

To create a statistic for the sequence, do the following:

> **select sequence(s)** | **Toolbox in the Menu Bar** | **General Sequence Analyses ( )** | **Create Sequence Statistics ( )**

This opens a dialog where you can alter your choice of sequences which you want to create statistics for. You can also add sequence lists.

Notice! You cannot create statistics for DNA and protein sequences at the same time.

When the sequences are selected, click **Next**.

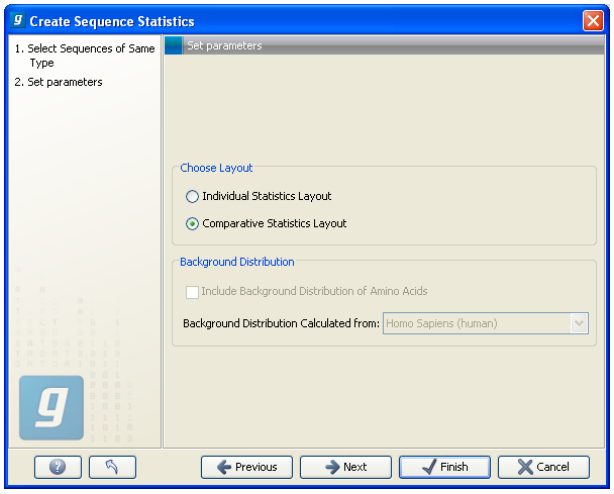


Figure 12.14: *Setting parameters for the sequence statistics.*

This opens the dialog displayed in figure 12.14.

The dialog offers to adjust the following parameters:

- **Individual statistics layout.** If more sequences were selected in **Step 1**, this function generates separate statistics for each sequence.
- **Comparative statistics layout.** If more sequences were selected in **Step 1**, this function generates statistics with comparisons between the sequences.

You can also choose to include Background distribution of amino acids. If this box is ticked, an extra column with amino acid distribution of the chosen species, is included in the table output. (The distributions are calculated from UniProt `www.uniprot.org` version 6.0, dated September 13 2005.)

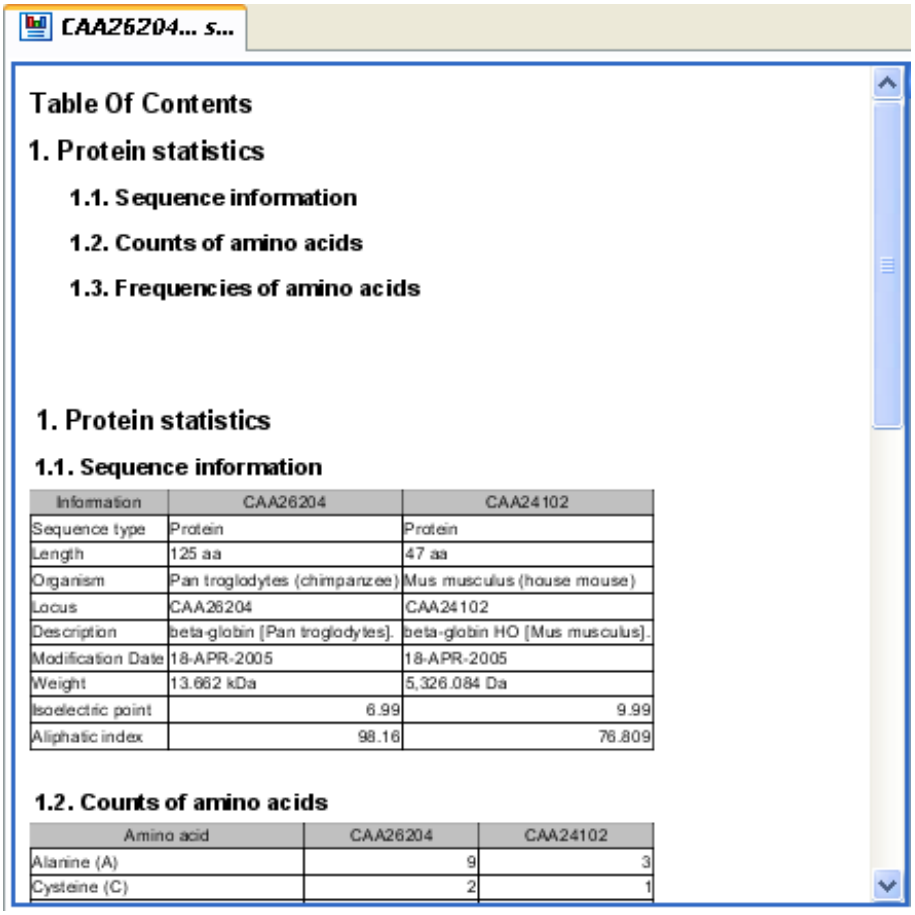
Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. An example of protein sequence statistics is shown in figure 12.15.

Nucleotide sequence statistics are generated using the same dialog as used for protein sequence statistics. However, the output of Nucleotide sequence statistics is less extensive than that of the protein sequence statistics.

Notice! The headings of the tables change depending on whether you calculate 'individual' or 'comparative' sequence statistics.

The output of comparative protein sequence statistics include:

- Sequence information:
  - Sequence type
  - Length
  - Organism
  - Locus
  - Description
  - Modification Date



Information	CAA26204	CAA24102
Sequence type	Protein	Protein
Length	125 aa	47 aa
Organism	Pan troglodytes (chimpanzee)	Mus musculus (house mouse)
Locus	CAA26204	CAA24102
Description	beta-globin [Pan troglodytes].	beta-globin HO [Mus musculus].
Modification Date	18-APR-2005	18-APR-2005
Weight	13.662 kDa	5,326.084 Da
Isoelectric point	6.99	9.99
Aliphatic index	98.16	76.809

Amino acid	CAA26204	CAA24102
Alanine (A)	9	3
Cysteine (C)	2	1

Figure 12.15: *Comparative sequence statistics.*

  - Weight
  - Isoelectric point
  - Aliphatic index

- Half-life
- Extinction coefficient
- Counts of Atoms
- Frequency of Atoms
- Count of hydrophobic and hydrophilic residues
- Frequencies of hydrophobic and hydrophilic residues
- Count of charged residues
- Frequencies of charged residues
- Amino acid distribution
- Histogram of amino acid distribution
- Annotation table

- Counts of di-peptides
- Frequency of di-peptides

The output of nucleotide sequence statistics include:

- General statistics:
  - Sequence type
  - Length
  - Organism
  - Locus
  - Description
  - Modification Date
  - Weight
- Atomic composition
- Nucleotide distribution table
- Nucleotide distribution histogram
- Annotation table
- Counts of di-nucleotides
- Frequency of di-nucleotides

A short description of the different areas of the statistical output is given in section 12.4.2.

12.4.1 Sequence statistics output

The entire statistical output can be printed. To do so, click the **Print** icon (( )).

12.4.2 Bioinformatics explained: Protein statistics

Every protein holds specific and individual features which are unique to that particular protein. Features such as isoelectric point or amino acid composition can reveal important information of a novel protein. Many of the features described below are calculated in a simple way.

Molecular weight

The molecular weight is the mass of a protein or molecule. The molecular weight is simply calculated as the sum of the atomic mass of all the atoms in the molecule.

The weight of a protein is usually represented in Daltons (Da).

A calculation of the molecular weight of a protein does not usually include additional posttranslational modifications. For native and unknown proteins it tends to be difficult to assess whether posttranslational modifications such as glycosylations are present on the protein, making a calculation based solely on the amino acid sequence inaccurate. The molecular weight can be determined very accurately by mass-spectrometry in a laboratory.

Isoelectric point

The isoelectric point (pI) of a protein is the pH where the proteins has no net charge. The pI is calculated from the pKa values for 20 different amino acids. At a pH below the pI, the protein carries a positive charge, whereas if the pH is above pI the proteins carry a negative charge. In other words, pI is high for basic proteins and low for acidic proteins. This information can be used in the laboratory when running electrophoretic gels. Here the proteins can be separated, based on their isoelectric point.

Aliphatic index

The aliphatic index of a protein is a measure of the relative volume occupied by aliphatic side chain of the following amino acids: alanine, valine, leucine and isoleucine. An increase in the aliphatic index increases the thermostability of globular proteins. The index is calculated by the following formula.

$$Aliphatic\ index = X(Ala) + a * X(Val) + b * X(Leu) + b * (X)Ile \tag{12.1}$$

X(Ala), *X(Val)*, *X(Ile)* and *X(Leu)* are the amino acid compositional fractions. The constants a and b are the relative volume of valine (a=2.9) and leucine/isoleucine (b=3.9) side chains compared to the side chain of alanine [Ikai, 1980].

Estimated half-life

The half life of a protein is the time it takes for the protein pool of that particular protein to be reduced to the half. The half life of proteins is highly dependent on the presence of the N-terminal amino acid, thus overall protein stability [Bachmair et al., 1986, Gonda et al., 1989, Tobias et al., 1991]. The importance of the N-terminal residues is generally known as the 'N-end rule'. The N-end rule and consequently the N-terminal amino acid, simply determines the half-life of proteins. The estimated half-life of proteins have been investigated in mammals, yeast and *E. coli* (see Table 12.2). If leucine is found N-terminally in mammalian proteins the estimated half-life is 5.5 hours.

Extinction coefficient

This measure indicates how much light is absorbed by a protein at a particular wavelength. The extinction coefficient is measured by UV spectrophotometry, but can also be calculated. The amino acid composition is important when calculating the extinction coefficient. The extinction coefficient is calculated from the absorbance of cysteine, tyrosine and tryptophan using the following equation:

$$Ext(Protein) = count(Cystine) * Ext(Cystine) + count(Tyr) * Ext(Tyr) + count(Trp) * Ext(Trp), \tag{12.2}$$

where Ext is the extinction coefficient of amino acid in question. At 280nm the extinction coefficients are: Cys=120, Tyr=1280 and Trp=5690.

This equation is only valid under the following conditions:

Amino acid	Mammalian	Yeast	E. coli
Ala (A)	4.4 hour	>20 hours	>10 hours
Cys (C)	1.2 hours	>20 hours	>10 hours
Asp (D)	1.1 hours	3 min	>10 hours
Glu (E)	1 hour	30 min	>10 hours
Phe (F)	1.1 hours	3 min	2 min
Gly (G)	30 hours	>20 hours	>10 hours
His (H)	3.5 hours	10 min	>10 hours
Ile (I)	20 hours	30 min	>10 hours
Lys (K)	1.3 hours	3 min	2 min
Leu (L)	5.5 hours	3 min	2 min
Met (M)	30 hours	>20 hours	>10 hours
Asn (N)	1.4 hours	3 min	>10 hours
Pro (P)	>20 hours	>20 hours	?
Gln (Q)	0.8 hour	10 min	>10 hours
Arg (R)	1 hour	2 min	2 min
Ser (S)	1.9 hours	>20 hours	>10 hours
Thr (T)	7.2 hours	>20 hours	>10 hours
Val (V)	100 hours	>20 hours	>10 hours
Trp (W)	2.8 hours	3 min	2 min
Tyr (Y)	2.8 hours	10 min	2 min

Table 12.2: **Estimated half life**. Half life of proteins where the N-terminal residue is listed in the first column and the half-life in the subsequent columns for mammals, yeast and *E. coli*.

- pH 6.5
- 6.0 M guanidium hydrochloride
- 0.02 M phosphate buffer

The extinction coefficient values of the three important amino acids at different wavelengths are found in [Gill and von Hippel, 1989].

Knowing the extinction coefficient, the absorbance (optical density) can be calculated using the following formula:

$$Absorbance(Protein) = \frac{Ext(Protein)}{Molecular\ weight} \tag{12.3}$$

Two values are reported. The first value is computed assuming that all cysteine residues appear as half cystines, meaning they form di-sulfide bridges to other cysteines. The second number assumes that no di-sulfide bonds are formed.

Atomic composition

Amino acids are indeed very simple compounds. All 20 amino acids consist of combinations of only five different atoms. The atoms which can be found in these simple structures are: Carbon, Nitrogen, Hydrogen, Sulfur, Oxygen. The atomic composition of a protein can for example be used to calculate the precise molecular weight of the entire protein.

Total number of negatively charged residues (Asp+Glu)

At neutral pH, the fraction of negatively charged residues provides information about the location of the protein. Intracellular proteins tend to have a higher fraction of negatively charged residues than extracellular proteins.

Total number of positively charged residues (Arg+Lys)

At neutral pH, nuclear proteins have a high relative percentage of positively charged amino acids. Nuclear proteins often bind to the negatively charged DNA, which may regulate gene expression or help to fold the DNA. Nuclear proteins often have a low percentage of aromatic residues [Andrade et al., 1998].

Amino acid distribution

Amino acids are the basic components of proteins. The amino acid distribution in a protein is simply the percentage of the different amino acids represented in a particular protein of interest. Amino acid composition is generally conserved through family-classes in different organisms which can be useful when studying a particular protein or enzymes across species borders. Another interesting observation is that amino acid composition variate slightly between proteins from different subcellular localizations. This fact has been used in several computational methods, used for prediction of subcellular localization.

Annotation table

This table provides an overview of all the different annotations associated with the sequence and their incidence.

Dipeptide distribution

This measure is simply a count, or frequency, of all the observed adjacent pairs of amino acids (dipeptides) found in the protein. It is only possible to report neighboring amino acids. Knowledge on dipeptide composition have previously been used for prediction of subcellular localization.

Creative Commons License
All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and "CLC bio" has to be clearly labelled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.

See http://creativecommons.org/licenses/by-nc-nd/2.5/ for more about how you may use the contents.

12.5 Join sequences

CLC Gene Workbench can join several nucleotide or protein sequences into one sequence. This feature can for example be used to construct "supergenes" for phylogenetic inference by joining several disjoint genes into one. Note, that when sequences are joined, all their annotations are carried over to the new spliced sequence.

Two (or more) sequences can be joined by:

select sequences to join | **Toolbox in the Menu Bar** | **General Sequence Analyses** | **Join sequences ( )**

or **select sequences to join** | **right-click either selected sequence** | **Toolbox** | **General Sequence Analyses** | **Join sequences ( )**

This opens the dialog shown in figure 12.16.

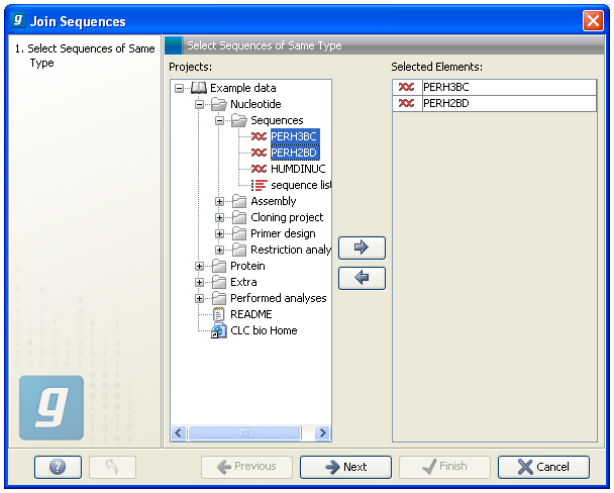


Figure 12.16: *Selecting two alignments to be joined.*

If you have selected some sequences before choosing the Toolbox action, they are now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences from the **Project Tree**. Click **Next** opens the dialog shown in figure 12.17.

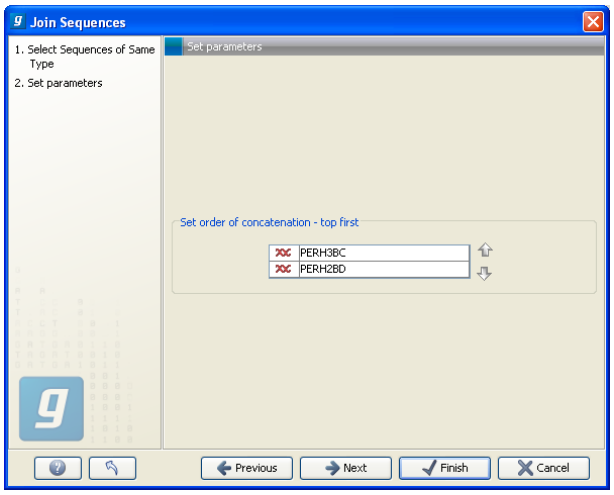


Figure 12.17: *Setting the order in which sequences are joined.*

In step 2 you can change the order in which the sequences will be joined. Select a sequence and

use the arrows to move the selected sequence up or down.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

The result is shown in figure 12.18.

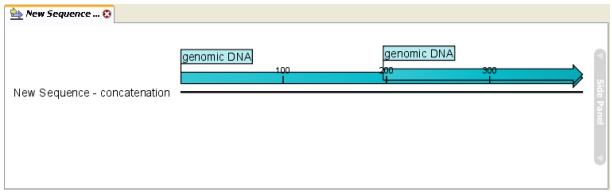


Figure 12.18: *The result of joining sequences is a new sequence containing all the annotations of the joined sequences.*

12.6 Motif Search

CLC Gene Workbench offers advanced and versatile options to search for unknown sequence patterns or known motifs represented either by a literal string or a regular expression. These advanced search capabilities are available for use in both DNA and protein sequences.

Difference between Motif Search and Pattern Discovery

In motif search (see 12.6), the user has some predefined knowledge about the pattern/motif of interest. This motif is defined by the user and the algorithm runs through the entire sequence and looks for identical or degenerate patterns. Motif search handles ambiguous characters in the way that two residues are different if they do not have any residues in common. For example: For nucleotides, N matches any character and R matches A, G. For proteins, X matches any character and Z matches E, Q.

Our pattern discovery algorithm (see 12.7) is based on proprietary hidden Markov models (HMM) and scans the entire sequence (one or more) for patterns which may be unknown to the user.

Motifs

If you have a known motif represented by a literal string or a sequence pattern of interest, you can search for them using the *CLC Gene Workbench*. Patterns and motifs can be searched with different levels of degeneracy in both DNA and protein sequences.

You can also search for matches with known motifs represented by a regular expression: A regular expressions is a string that describes or matches a set of strings, according to certain syntax rules. They are usually used to give a concise description of a set, without having to list all elements. The simplest form of a regular expression is a literal string. You are limited to the following syntax rules (See the Java regular expression syntax):

> $[A - Z]$ will match the characters A through Z (Range). You can also put single characters between the brackets: The expression $[AGT]$ matches the characters A, G or T.
>
> $[A - D[M - P]]$ will match the characters A through D and M through P (Union). You can also put single characters between the brackets: The expression $[AG[M - P]]$ matches the characters A, G and M through P.

[$A-M$&&[$H-P$]] will match the characters between A and M lying between H and P (Intersection). You can also put single characters between the brackets. The expression [$A-M$&&[$HGTDA$]] matches the characters A through M which is H, G, T, D or A.

[^$A-M$] will match any character except those between A and M (Excluding). You can also put single characters between the brackets: The expression [^AG] matches any character except A and G.

[$A-Z$&&[^$M-P$]] will match any character A through Z except those between M and P (Subtraction). You can also put single characters between the brackets: The expression [$A-P$&&[^CG]] matches any character between A and P except C and G.

The symbol . matches any character.

$X\{n\}$ will match a repetition of an element indicated by following that element with a numerical value or a numerical range between the curly brackets. For example, $ACG\{2\}$ matches the string $ACGACG$.

$X\{n,m\}$ will match a certain number of repetitions of an element indicated by following that element with two numerical values between the curly brackets. The first number is a lower limit on the number of repetitions and the second number is an upper limit on the number of repetitions. For example, $ACT\{1,3\}$ matches $ACT, ACTACT$ and $ACTACTACT$.

$X\{n,\}$ represents a repetition of an element at least n times. For example, $AC\{2,\}$ matches all strings $ACAC$, $ACACAC$, $ACACACAC$,...

The symbol ^ restricts the search to the beginning of your sequence. For example, if you search through a sequence with the regular expression ^AC, the algorithm will find a match if AC occurs in the beginning of the sequence.

The symbol $ restricts the search to the end of your sequence. For example, if you search through a sequence with the regular expression GT\$, the algorithm will find a match if GT occurs in the end of the sequence.

Examples The expression $[ACG][^AC]G\{2\}$ matches all strings of length 4, where the first character is A, C or G and the second is any character except A, C and the third and fourth character is G. The expression $G.[^A]\$$ matches all strings of length 3 in the end of your sequence, where the first character is C, the second any character and the third any character except A.

For proteins, you can enter different protein patterns from the PROSITE database (protein patterns using regular expressions and describing specific amino acid sequences). The PROSITE database contains a great number of patterns and have been used to identify related proteins.

In order to search for a known motif:

Select DNA or protein sequence(s) | **Toolbox in the Menu Bar** | **General Sequence Analyses ( )** | **Motif Search ( )**

or **Right-click DNA or protein sequence(s)** | **Toolbox** | **General Sequence Analyses ( )** | **Motif Search ( )**

If a sequence was selected before choosing the Toolbox action, the sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the **Project Tree**.

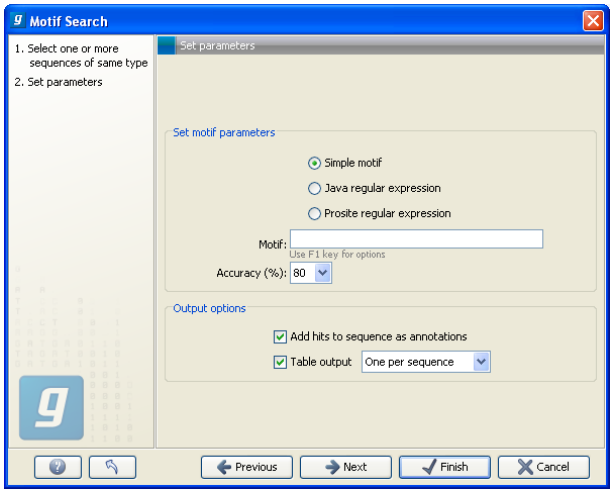


Figure 12.19: *Setting parameters for the motif search. See text for details.*

You can perform the analysis on several DNA or several protein sequences at a time. If the analysis is performed on several sequences at a time the method will search for patterns in the sequences and open a new view for each of the sequences.

Click **Next** to adjust parameters (see figure 12.19).

12.6.1 Motif search parameter settings

Various parameters can be set prior to the motif search. The parameters are listed below and a screen shot of the parameter settings can be seen in figure 12.19.

- **Motif types**
  You can choose literal string (simple motif) or Java regular expression as your motif type. For proteins, you can choose to search with a Prosite regular expression.
- **Motif**
  If you choose to search with a simple motif, you should enter a literal string as your motif. Ambiguous amino acids and nucleotides are allowed. Example; ATGATGNNATG. If your motif type is Java regular expression, you should enter a regular expression according to the syntax rules above. Press **F1** key for options. For proteins, you can search with a Prosite regular expression and you should enter a protein pattern from the PROSITE database.
- **Accuracy**
  If you search with a simple motif, you can adjust the accuracy of the search string to the match on the sequence.
- **Table output**
  Opens the motifs or patterns found in a table view. It is possible to see one table per sequence but it is also possible with one table for multiple sequences.
- **Add motif to sequence as annotation**
  Check this box to add search strings found as annotations on the sequence.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. This will open a view showing the motifs or patterns found as annotations on the original

sequence (see figure 12.20). If you have selected several sequences, a corresponding number of views will be opened.

Figure 12.20: *Sequence view displaying the pattern found. The search string was 'QRQXRXXXXQQ'.*

12.6.2 Motif search output

If the analysis is performed on several sequences at a time the method will search for patterns in the sequences and open a new view for each of the sequences. If wanted, annotations on patterns found can be added to all the sequences. Each pattern found will be represented as an annotation of the type **Region**. More information on each motif or pattern found is available through the tooltip, including detailed information on the position of the pattern and how similar it was to the search string.

It is also possible to get a tabular view of all motifs or patterns found in either one combined table or in individual tables if multiple sequences were selected. Then each pattern found will be represented with its position in the sequence and the obtained accuracy score.

12.7 Pattern Discovery

With *CLC Gene Workbench* you can perform pattern discovery on both DNA and protein sequences. Advanced hidden Markov models can help to identify unknown sequence patterns across single or even multiple sequences.

In order to search for unknown patterns:

Select DNA or protein sequence(s) | **Toolbox in the Menu Bar** | **General Sequence Analyses** ( ) | **Pattern Discovery** ( )

or **right-click DNA or protein sequence(s)** | **Toolbox** | **General Sequence Analyses** ( ) | **Pattern Discovery** ( )

If a sequence was selected before choosing the Toolbox action, the sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the **Project Tree**.

You can perform the analysis on several DNA or several protein sequences at a time. If the analysis is performed on several sequences at a time the method will search for patterns which is common between all the sequences. Annotations will be added to all the sequences and a view is opened for each sequence.

Click **Next** to adjust parameters (see figure 12.21).

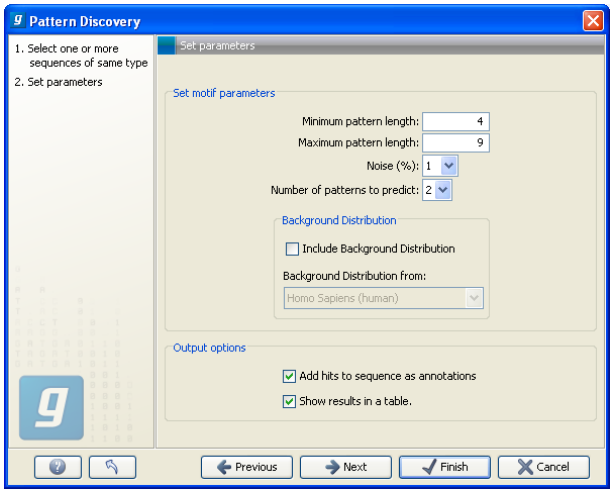


Figure 12.21: *Setting parameters for the pattern discovery. See text for details.*

12.7.1 Pattern discovery search parameters

Various parameters can be set prior to the pattern discovery. The parameters are listed below and a screen shot of the parameter settings can be seen in figure 12.21.

- **Minimum pattern length**
  Here, the minimum length of patterns to search for, can be specified.
- **Maximum pattern length**
  Here, the maximum length of patterns to search for, can be specified.
- **Noise (%)**
  Specify noise-level of the model. This parameter has influence on the level of degeneracy of patterns in the sequence(s). The noise parameter can be 1,2,5 or 10 percent.
- **Number of different kinds of patterns to predict**
  Number of iterations the algorithm goes through. After the first iteration, we force predicted pattern-positions in the first run to be member of the background: In that way, the algorithm finds new patterns in the second iteration. Patterns marked 'Pattern1' have the highest confidence. The maximal iterations to go through is 3.
- **Show result of patterns discovery in a table**
  Generate a tabular output which displays patterns found.
- **Include Background Distribution of Amino Acids**
  For protein sequences it is possible to include information on the background distribution of amino acids from a range of organisms.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. This will open a view showing the patterns found as annotations on the original sequence (see figure 12.22). If you have selected several sequences, a corresponding number of views will be opened.

Figure 12.22: *Sequence view displaying two discovered patterns.*

12.7.2 Pattern search output

If the analysis is performed on several sequences at a time the method will search for patterns in the sequences and open a new view for each of the sequences, in which a pattern was discovered. Each novel pattern will be represented as an annotation of the type **Region**. More information on each found pattern is available through the tooltip, including detailed information on the position of the pattern and quality scores.

It is also possible to get a tabular view of all found patterns in one combined table. Then each found pattern will be represented with various information on obtained scores, quality of the pattern and position in the sequence.

Chapter 13

Nucleotide analyses

Contents

13.1 Convert DNA to RNA **163**
13.2 Convert RNA to DNA **164**
13.3 Reverse complements of sequences **165**
13.4 Translation of DNA or RNA to protein **166**
13.4.1 Translate part of a nucleotide sequence 167
13.5 Annotate with SNPs **167**
13.6 Find open reading frames **168**
13.6.1 Open reading frame parameters 168

CLC Gene Workbench 2.0 offers different kinds of sequence analyses, which only apply to DNA and RNA.

13.1 Convert DNA to RNA

CLC Gene Workbench 2.0 lets you convert a DNA sequence into RNA, substituting the T residues (Thymine) for U residues (Urasil):

select a DNA sequence in the Navigation Area | Toolbox in the Menu Bar | Nucleotide Analyses ( ) | Convert DNA to RNA ( )

or **right-click a sequence in Navigation Area | Toolbox | Nucleotide Analyses ( ) | Convert DNA to RNA ( )**

This opens the dialog displayed in figure 13.1:

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the **Project Tree**.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

Notice! You can select multiple DNA sequences and sequence lists at a time. If the sequence list contains RNA sequences as well, they will not be converted.

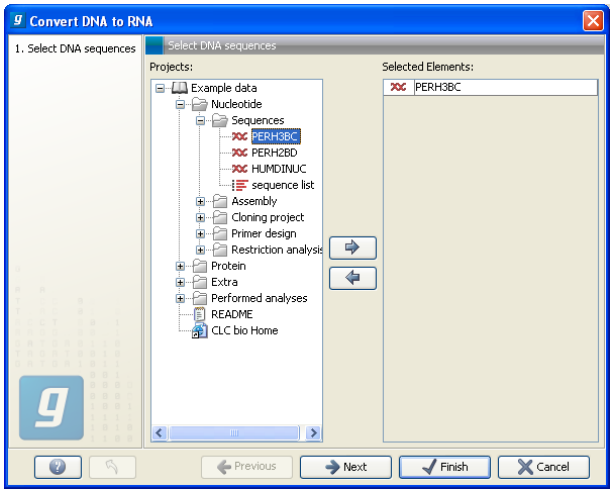


Figure 13.1: *Translating DNA to RNA.*

13.2 Convert RNA to DNA

CLC Gene Workbench 2.0 lets you convert an RNA sequence into DNA, substituting the U residues (Urasil) for T residues (Thymine):

select an RNA sequence in the Navigation Area | Toolbox in the Menu Bar | Nucleotide Analyses ( ) | Convert RNA to DNA ( )

or **right-click a sequence in Navigation Area | Toolbox | Nucleotide Analyses ( ) | Convert RNA to DNA ( )**

This opens the dialog displayed in figure 13.2:

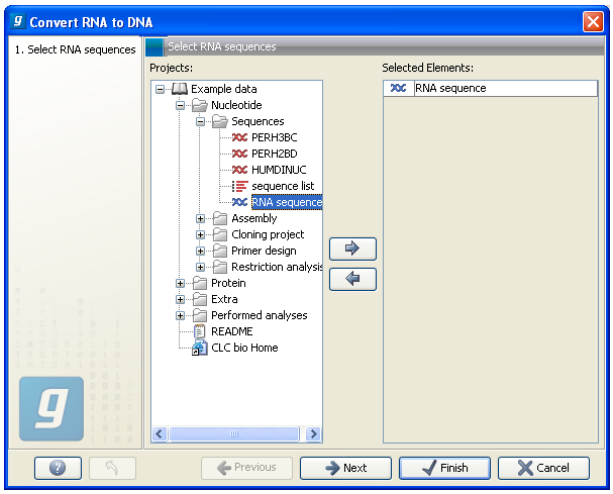


Figure 13.2: *Translating RNA to DNA.*

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the **Project Tree**.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

This will open a new view in the **View Area** displaying the new DNA sequence. The new sequence is not saved automatically. To save the protein sequence, drag it into the **Navigation Area** or

press Ctrl + S (⌘ + S on Mac) to activate a save dialog.

Notice! You can select multiple RNA sequences and sequence lists at a time. If the sequence list contains DNA sequences as well, they will not be converted.

13.3 Reverse complements of sequences

CLC Gene Workbench 2.0 is able to create the reverse complement of a nucletide sequence. By doing that, a new sequence is created which also has all the annotations reversed since they now occupy the opposite strand of their previous location.

To quickly obtain the reverse complement of a sequence or part of a sequence, you may select a region on the negative strand and open it in a new view:

right-click a selection on the negative strand | Open selection in a new view

By doing that, the sequence will be reversed. This is only possible when the double stranded view option is enabled. It is possible to copy the selection and paste it in a word processing program or an e-mail. To obtain a reverse complement of an entire sequence:

select a sequence in the Navigation Area | Toolbox in the Menu Bar | Nucleotide Analyses ( ) | Create Reverse Complement ( )

or **right-click a sequence in Navigation Area | Toolbox | Nucleotide Analyses () | Create Reverse Complement ( )**

This opens the dialog displayed in figure 13.3:

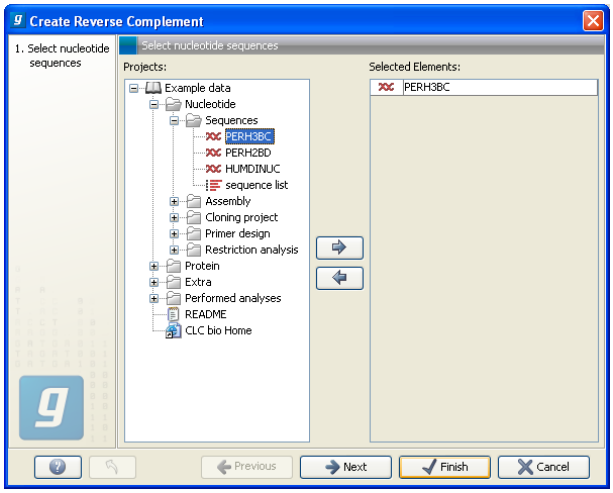


Figure 13.3: *Creating a reverse complement sequence.*

If a sequence was selected before choosing the Toolbox action, the sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the **Project Tree**.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

This will open a new view in the **View Area** displaying the reverse complement of the selected sequence. The new sequence is not saved automatically. To save the protein sequence, drag it into the **Navigation Area** or press Ctrl + S (⌘ + S on Mac) to activate a save dialog.

13.4 Translation of DNA or RNA to protein

In *CLC Gene Workbench 2.0* you can translate a nucleotide sequence into a protein sequence using the **Toolbox** tools. Usually, you use the +1 reading frame which means that the translation starts from the first nucleotide. Stop codons result in an asterisk being inserted in the protein sequence at the corresponding position. It is possible to translate in any combination of the six reading frames in one analysis. To translate:

select a nucleotide sequence | **Toolbox in the Menu Bar** | **Nucleotide Analyses ( )** | **Translate to Protein ( )**

or **right-click a nucleotide sequence** | **Toolbox** | **Nucleotide Analyses ( )** | **Translate to Protein ( )**

This opens the dialog displayed in figure 13.4:

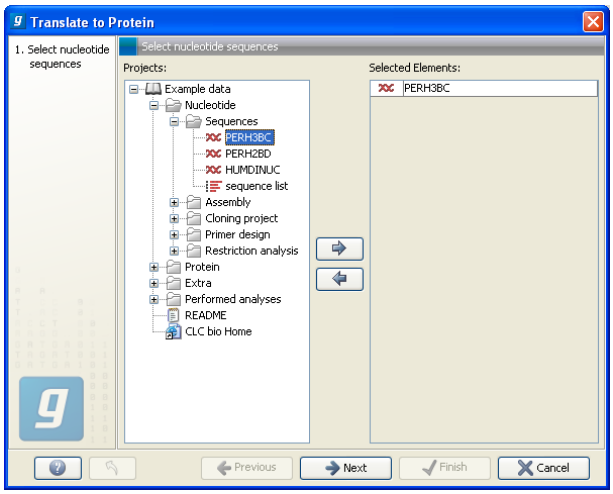


Figure 13.4: *Choosing sequences for translation.*

If a sequence was selected before choosing the Toolbox action, the sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the **Project Tree**.

Click **Next** to set reading frames, select if you want to translate all coding regions of the sequence and choose translation tables. Clicking **Next** generates the dialog seen in figure 13.5:

The translation tables in *CLC Gene Workbench* are updated regularly from NCBI. Therefore the tables are not available in this printable version of the user manual. Instead the tables are included in the **Help**-menu in the **Menu Bar** under **Background Information**.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. The newly created protein is shown, but is not saved automatically.

There are also new views of proteins for every **CDS** or **ORF** annotation if you have selected to translate all coding regions.

To save a protein sequence, drag it into the **Navigation Area** or press Ctrl + S (⌘ + S on Mac) to activate a save dialog.

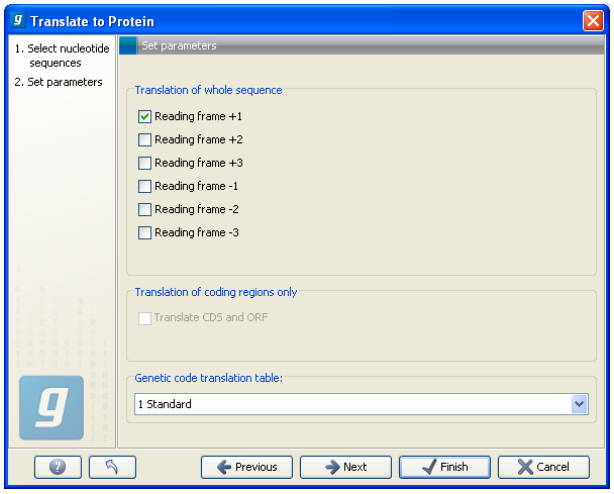


Figure 13.5: *Choosing +1 and +3 reading frames, and the standard translation table.*

13.4.1 Translate part of a nucleotide sequence

If you want to make separate translations of *all* the coding regions of a nucleotide sequence, you can check the option: "Translate CDS and ORF" in the translation dialog (see figure 13.5).

If you want to translate a *specific* coding region, which is annotated on the sequence, use the following procedure:

Open the nucleotide sequence | **right-click the ORF or CDS annotation** | **Translate CDS/ORF** | **choose a translation table** | **OK**

If the annotation contains information about the translation, this information will be used, and you do not have to specify a translation table.

The CDS and ORF annotations are colored yellow as default.

13.5 Annotate with SNPs

CLC Gene Workbench 2.0 can annotate sequences with Single Nucleotide Polymorphism (SNP) as found in the NCBI online dbSNP database. A SNP is a mutation of a single nucleotide in a DNA sequence.

To annotate with SNPs:

select a nucleotide sequence | **Toolbox in the Menu Bar** | **Nucleotide Analyses** | **Annotate with SNPs from NCBI**

or **right-click a nucleotide sequence** | **Toolbox** | **Nucleotide Analyses** | **Annotate with SNPs from NCBI**

If a sequence was selected before choosing the Toolbox action, the sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the **Project Tree**. When you have selected the desired sequences, click finish. Warning: annotating a sequence with SNPs may take a while, depending on sequence size and the availability of the NCBI server. Furthermore, only sequences which have a GenBank accession number can be annotated with SNPs.

After executing the action you will see the sequence annotated with the SNPs as in figure 13.6:

Figure 13.6: *A sequence annotated with SNP's.*

13.6 Find open reading frames

CLC Gene Workbench 2.0 has a basic functionality for gene finding in the form of open reading frame (ORF) determination. The ORFs will be shown as annotations on the sequence. You have the option of choosing translation table, start codons, minimum length and other parameters for finding the ORFs. These parameters will be explained in this section.

To find open reading frames:

select a nucleotide sequence | **Toolbox in the Menu Bar** | **Nucleotide Analyses ( )** | **Find Open Reading Frames ( )**

or **right-click a nucleotide sequence** | **Toolbox** | **Nucleotide Analyses ( )** | **Find Open Reading Frames ( )**

This opens the dialog displayed in figure 13.7:

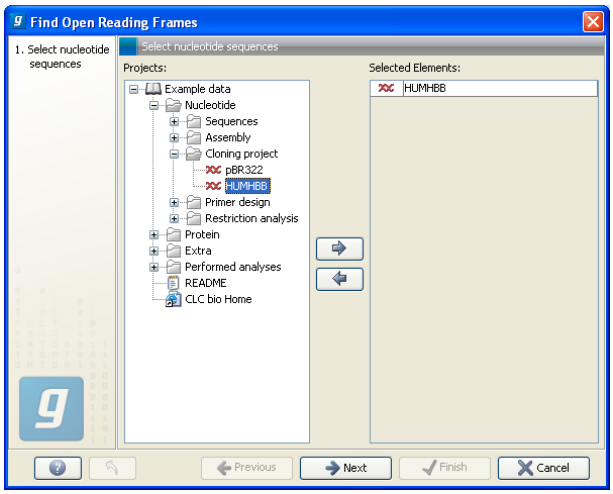


Figure 13.7: *Create Reading Frame dialog.*

If a sequence was selected before choosing the Toolbox action, the sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the **Project Tree**.

If you want to adjust the parameters for finding open reading frames click **Next**.

13.6.1 Open reading frame parameters

This opens the dialog displayed in figure 13.8:

The adjustable parameters for the search are:

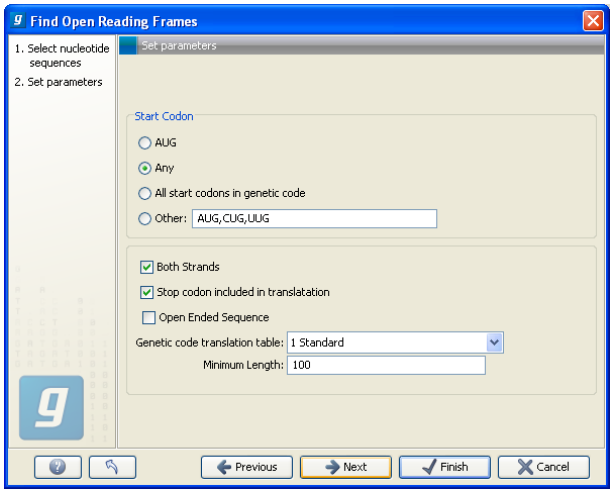


Figure 13.8: *Create Reading Frame dialog.*

- **Start Codon**:
  - **AUG**. Most commonly used start codon.
  - **Any**.
  - **All start codons in genetic code**.
  - **Other**. Here you can specify a number of start codons separated by commas.
- **Both Strands**. Finds reading frames on both strands.
- **Stop Codon included in Annotation** The ORFs will be shown as annotations which can include the stop codon if this option is checked.
- **Open Ended Sequence**. Allows the ORF to start or end outside the sequence. If the sequence studied is a part of a larger sequence, it may be advantageous to allow the ORF to start or end outside the sequence.
- **Genetic code translation table**. The translation tables are occasionally updated from NCBI. The tables are not available in this printable version of the user manual. Instead, the tables are included in the **Help**-menu in the **Menu Bar** under **Background Information**.
- **Minimum Length**. Specifies the minimum length for the ORFs to be found.

Using open reading frames for gene finding is a fairly simple approach which is likely to predict genes which are not real. Setting a relatively high minimum length of the ORFs will reduce the number of false positive predictions, but at the same time short genes may be missed (see figure 13.9).

Finding open reading frames is often a good first step in annotating sequences such as cloning vectors or bacterial genomes. For eukaryotic genes, ORF determination may not always be very helpful since the intron/exon structure is not part of the algorithm.

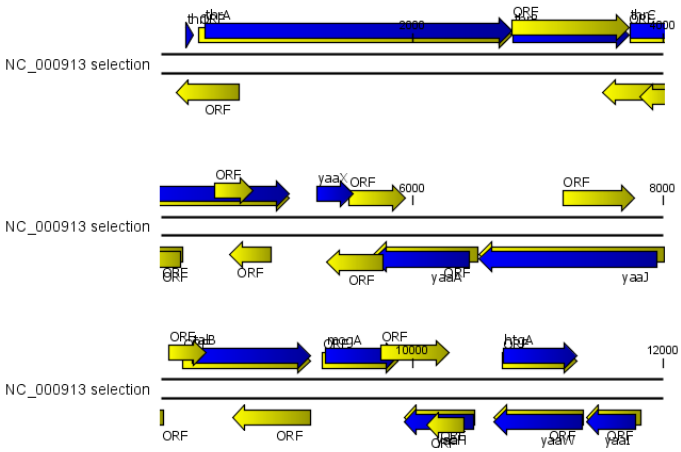


Figure 13.9: *The first 12,000 positions of the* E. coli *sequence NC_000913 downloaded from GenBank. The blue (dark) annotations are the genes while the yellow (brighter) annotations are the ORFs with a length of at least 100 amino acids. On the positive strand around position 11,000, a gene starts before the ORF. This is due to the use of the standard genetic code rather than the bacterial code. This particular gene starts with CTG, which is a start codon in bacteria. Two short genes are entirely missing, while a handful of open reading frames do not correspond to any of the annotated genes.*

Chapter 14

Protein analyses

Contents

14.1 Protein charge 171
14.1.1 Modifying the layout 172
14.2 Hydrophobicity 174
14.2.1 Hydrophobicity plot 174
14.2.2 Hydrophobicity graphs along sequence 176
14.2.3 Bioinformatics explained: Protein hydrophobicity 177
14.3 Reverse translation from protein into DNA 179
14.3.1 Reverse translation parameters 179
14.3.2 Bioinformatics explained: Reverse translation 181

CLC Gene Workbench 2.0 offers analyses of proteins as described in this chapter.

14.1 Protein charge

In *CLC Gene Workbench* you can create a graph in the electric charge of a protein as a function of pH. This is particularly useful for finding the net charge of the protein at a given pH. This knowledge can be used e.g. in relation to isoelectric focusing on the first dimension of 2D-gel electrophoresis. The isoelectric point (pI) is found where the net charge of the protein is zero. The calculation of the protein charge does not include knowledge about any potential post-translational modifications the protein may have.

In order to calculate the protein charge:

Select a protein sequence | **Toolbox in the Menu Bar** | **Protein Analyses ( )** | **Create Protein Charge Plot ( )**

or **right-click a protein sequence** | **Toolbox** | **Protein Analyses ( )** | **Create Protein Charge Plot ( )**

This opens the dialog displayed in figure 14.1:

If a sequence was selected before choosing the Toolbox action, the sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or

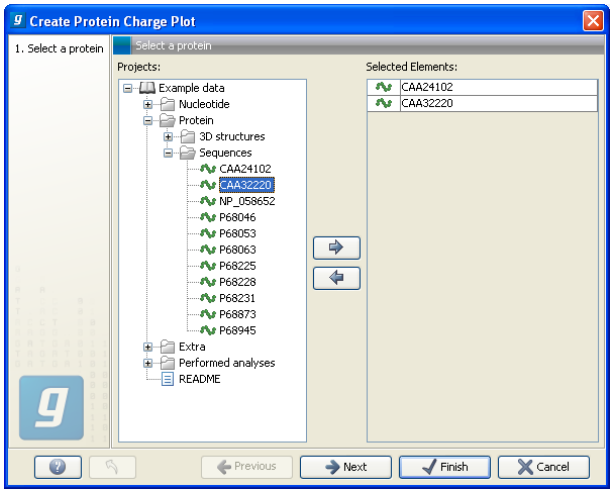


Figure 14.1: *Choosing protein sequences to calculate protein charge.*

sequence lists from the **Project Tree**.

You can perform the analysis on several protein sequences at a time. This will result in one output graph showing protein charge graphs for the individual proteins.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

14.1.1 Modifying the layout

Figure 14.2 shows the electrical charges for three proteins. In the **Side Panel** to the right, you can modify the layout of the graph.

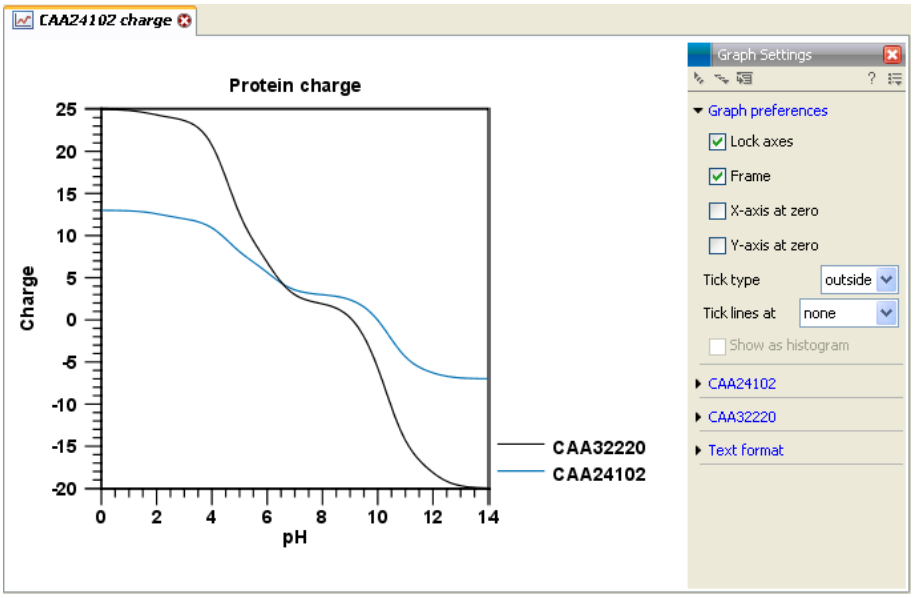


Figure 14.2: *View of the protein charge.*

Graph preferences

The **Graph preferences** apply to the whole graph:

- **Lock axis.**
  This will always show the axis even though the plot is zoomed to a detailed level.
- **Frame.**
  Toggles the frame of the graph.
- **X-axis at zero.**
  Toggles the x-axis at zero.
- **Y-axis at zero.**
  Toggles the y-axis at zero.
- **Tick type**
  - outside
  - inside
- **Tick lines at.**
  Shows a grid behind the graph.
  - none
  - major ticks
- **Show as histogram**
  For some data-series it is possible to see it as a histogram rather than a line plot.

Preferences for each protein

Underneath the **Graph preferences** you will find is a set of preferences for each protein in the graph. These preferences only apply to the curve for the specific protein.

- **Dot type**
  - none
  - cross
  - plus
  - square
  - diamond
  - circle
  - triangle
  - reverse triangle
  - dot
- **Dot color.** Allows you to choose between many different colors.
- **Line width**
  - thin
  - medium

  - wide

- **Line type**
  - none
  - line
  - long dash
  - short dash

- **Line color.** Allows you to choose between many different colors.

These settings will apply to both the curve and the legend.

Modifying labels and legends

Click the title of the graph, the axis-titles or the legend to edit the text.

14.2 Hydrophobicity

CLC Gene Workbench can calculate the hydrophobicity of protein sequences in different ways, using different algorithms. (See section 14.2.3). Furthermore, hydrophobicity of sequences can be displayed as hydrophobicity plots and as graphs along sequences. In addition, *CLC Gene Workbench 2.0* can calculate hydrophobicity for several sequences at the same time, and for alignments.

14.2.1 Hydrophobicity plot

To display the hydrophobicity for a protein sequence in a plot is done in the following way:

> **select a protein sequence in Navigation Area | Toolbox in the Menu Bar | Protein Analyses ( ) | Create Hydrophobicity Plot ( )**

This opens a dialog. The first step allows you to add or remove sequences. Clicking **Next** takes you through to **Step 2**, which is displayed in figure 14.3.

The **Window size** is the width of the window where the hydrophobicity is calculated. The wider the window, the less volatile the graph. You can chose from a number of hydrophobicity scales which are further explained in section 14.2.3 Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. The result can be seen in figure 14.4.

In *CLC Gene Workbench 2.0* it is possible to change the layout of the hydrophobicity plot through the **Side Panel**. The drop down menus are opened by clicking the black triangular arrows.

There are two kinds of view preferences. The graph preferences and preferences for the kind of hydrophobicity scale used to calculate the graph, e.g. Kyte-Doolittle.

The **Graph preferences** include:

- **Lock axis.**
  This will always show the axis even though the plot is zoomed to a detailed level.

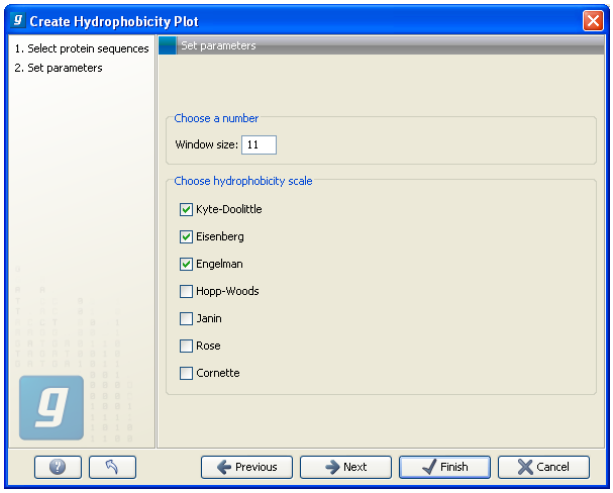


Figure 14.3: *Step two in the Hydrophobicity Plot allows you to choose hydrophobicity scale and the window size.*

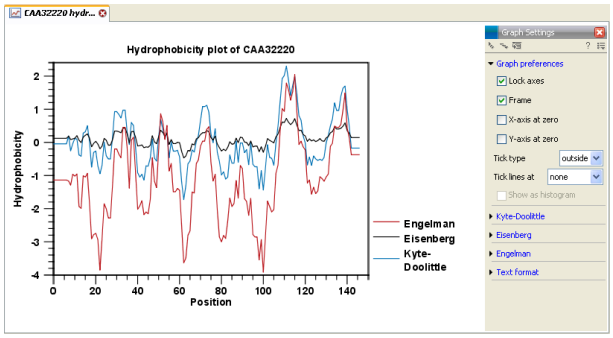


Figure 14.4: *The result of the hydrophobicity plot calculation and the associated Side Panel.*

- **Frame.**
  Toggles the frame of the graph.
- **X-axis at zero.**
  Toggles the x-axis at zero.
- **Y-axis at zero.**
  Toggles the y-axis at zero.
- **Tick type**
  - outside
  - inside
- **Tick lines at.**
  Shows a grid behind the graph.
  - none
  - major ticks
- **Show as histogram**
  For some data-series it is possible to see it as a histogram rather than a line plot.

The preferences for the different scales are identical and include the following:

- **Dot type.**
  Lets you choose the marking of dots in the graph.
- **Dot color.**
  Lets you choose the color of the dots.
- **Line width.**
  Applies to the line connecting the dots.
- **Line type.**
  Applies to the line connecting the dots.
- **Line color.**
  Applies to the line connecting the dots.

14.2.2 Hydrophobicity graphs along sequence

Hydrophobicity graphs along sequence can be displayed easily by activating the calculations from the **Side Panel** for a sequence.

right-click protein sequence in Navigation Area | Show | Sequence | open Hydrophobicity info in Side Panel

or **double-click protein sequence in Navigation Area | Show | Sequence | open Hydrophobicity info in Side Panel**

These actions result in the view displayed in figure 14.5.

Figure 14.5: *The different available scales in Hydrophobicity info in* **CLC Gene Workbench 2.0**.

The level of hydrophobicity is calculated on the basis of the different scales. The different scales add different values to each type of amino acid. The hydrophobicity score is then calculated as the sum of the values in a 'window', which is a particular range of the sequence. The window length can be set from 5 to 25 residues. The wider the window, the less fluctuations in the hydrophobicity scores. (For more about the theory behind hydrophobicity, see 14.2.3).

In the following we will focus on the different ways that *CLC Gene Workbench 2.0* offers to display the hydrophobicity scores. We use Kyte-Doolittle to explain the display of the scores, but the different options are the same for all the scales. Initially there are three options for displaying the hydrophobicity scores. You can choose one, two or all three options by selecting the boxes. (See figure 14.6).

Coloring the letters and their background. When choosing coloring of letters or coloring of their background, the color red is used to indicate high scores of hydrophobicity. A 'color-slider' allows you to amplify the scores, thereby emphasizing areas with high (or low, blue) levels of hydrophobicity. The color settings mentioned are default settings. By clicking the color bar just below the color slider you get the option of changing color settings.

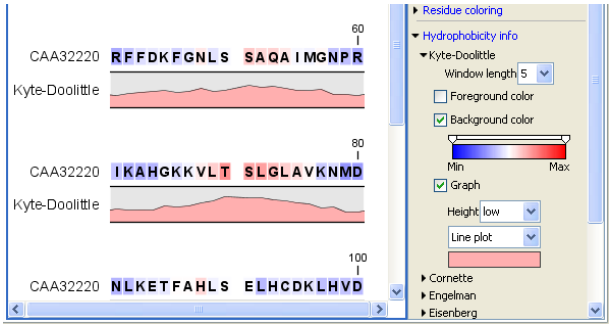


Figure 14.6: *The different ways of displaying the hydrophobicity scores, using the Kyte-Doolittle scale.*

Graphs along sequences. When selecting graphs, you choose to display the hydrophobicity scores underneath the sequence. This can be done either by a line-plot or bar-plot, or by coloring. The latter option offers you the same possibilities of amplifying the scores as applies for coloring of letters. The different ways to display the scores when choosing 'graphs' are displayed in figure 14.6. Notice that you can choose the height of the graphs underneath the sequence.

14.2.3 Bioinformatics explained: Protein hydrophobicity

Calculation of hydrophobicity is important to the identification of various protein features. This can be membrane spanning regions, antigenic sites, exposed loops or buried residues. Usually, these calculations are shown as a plot along the protein sequence, making it easy to identify the location of potential protein features.

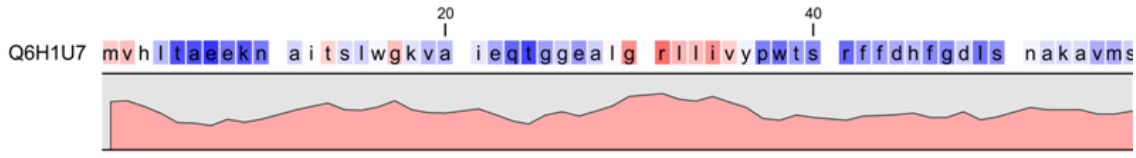


Figure 14.7: *Plot of hydrophobicity along the amino acid sequence. Hydrophobic regions on the sequence have higher numbers according to the graph below the sequence, furthermore hydrophobic regions are colored on the sequence. Red indicates regions with high hydrophobicity and blue indicates regions with low hydrophobicity.*

The hydrophobicity is calculated by sliding a fixed size window (of an odd number) over the protein sequence. At the central position of the window, the average hydrophobicity of the entire window is plotted (see figure 14.7).

Hydrophobicity scales

Several hydrophobicity scales have been published for various uses. Many of the commonly used hydrophobicity scales are described below.

Kyte-Doolittle scale. The Kyte-Doolittle scale is widely used for detecting hydrophobic regions in proteins. Regions with a positive value are hydrophobic. This scale can be used for identifying both surface-exposed regions as well as transmembrane regions, depending on the window size

used. Short window sizes of 5-7 generally work well for predicting putative surface-exposed regions. Large window sizes of 19-21 are well suited for finding transmembrane domains if the values calculated are above 1.6 [Kyte and Doolittle, 1982]. These values should be used as a rule of thumb and deviations from the rule may occur.

Engelman scale. The Engelman hydrophobicity scale, also known as the GES-scale, is another scale which can be used for prediction of protein hydrophobicity [Engelman et al., 1986]. As the Kyte-Doolittle scale, this scale is useful for predicting transmembrane regions in proteins.

Eisenberg scale. The Eisenberg scale is a normalized consensus hydrophobicity scale which shares many features with the other hydrophobocity scales [Eisenberg et al., 1984].

Hopp-Woods scale. Hopp and Woods developed their hydrophobicity scale for identification of potentially antigenic sites in proteins. This scale is basically a hydrophilic index where apolar residues have been assigned negative values. Antigenic sites are likely to be predicted when using a window size of 7 [Hopp and Woods, 1983].

Cornette scale. Cornette *et al.* computed an optimal hydrophobicity scale based on 28 published scales [Cornette et al., 1987]. This optimized scale is also suitable for prediction of alpha-helices in proteins.

Rose scale. The hydrophobicity scale by Rose *et al.* is correlated to the average area of buried amino acids in globular proteins [Rose et al., 1985]. This results in a scale which is not showing the helices of a protein, but rather the surface accessibility.

Janin scale. This scale also provides information about the accessible and buried amino acid residues of globular proteins [Janin, 1979].

Many more scales have been published throughout the last three decades. Even though more advanced methods have been developed for prediction of membrane spanning regions, the simple and very fast calculations are still highly used.

Other useful resources

AAindex: Amino acid index database
`http://www.genome.ad.jp/dbget/aaindex.html`

Creative Commons License
All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and "CLC bio" has to be clearly labelled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.

See `http://creativecommons.org/licenses/by-nc-nd/2.5/` for more about how you may use the contents.

aa	aa	Kyte-Doolittle	Hopp-Woods	Cornette	Eisenberg	Rose	Janin	Engelman (GES)
A	Alanine	1.80	-0.50	0.20	0.62	0.74	0.30	1.60
C	Cysteine	2.50	-1.00	4.10	0.29	0.91	0.90	2.00
D	Aspartic acid	-3.50	3.00	-3.10	-0.90	0.62	-0.60	-9.20
E	Glutamic acid	-3.50	3.00	-1.80	-0.74	0.62	-0.70	-8.20
F	Phenylalanine	2.80	-2.50	4.40	1.19	0.88	0.50	3.70
G	Glycine	-0.40	0.00	0.00	0.48	0.72	0.30	1.00
H	Histidine	-3.20	-0.50	0.50	-0.40	0.78	-0.10	-3.00
I	Isoleucine	4.50	-1.80	4.80	1.38	0.88	0.70	3.10
K	Lysine	-3.90	3.00	-3.10	-1.50	0.52	-1.80	-8.80
L	Leucine	3.80	-1.80	5.70	1.06	0.85	0.50	2.80
M	Methionine	1.90	-1.30	4.20	0.64	0.85	0.40	3.40
N	Asparagine	-3.50	0.20	-0.50	-0.78	0.63	-0.50	-4.80
P	Proline	-1.60	0.00	-2.20	0.12	0.64	-0.30	-0.20
Q	Glutamine	-3.50	0.20	-2.80	-0.85	0.62	-0.70	-4.10
R	Arginine	-4.50	3.00	1.40	-2.53	0.64	-1.40	-12.3
S	Serine	-0.80	0.30	-0.50	-0.18	0.66	-0.10	0.60
T	Threonine	-0.70	-0.40	-1.90	-0.05	0.70	-0.20	1.20
V	Valine	4.20	-1.50	4.70	1.08	0.86	0.60	2.60
W	Tryptophan	-0.90	-3.40	1.00	0.81	0.85	0.30	1.90
Y	Tyrosine	-1.30	-2.30	3.20	0.26	0.76	-0.40	-0.70

Table 14.1: *Hydrophobicity scales. This table shows seven different hydrophobicity scales which are generally used for prediction of e.g. transmembrane regions and antigenicity.*

14.3 Reverse translation from protein into DNA

A protein sequence can be back-translated into DNA using *CLC Gene Workbench 2.0*. Due to degeneracy of the genetic code every amino acid could translate into several different codons (only 20 amino acids but 64 different codons). Thus, the program offers a number of choices for determining which codons should be used. These choices are explained in this section.

In order to make a reverse translation:

Select a protein sequence | **Toolbox in the Menu Bar** | **Protein Analyses ( )** | **Reverse Translate ( )**

or **right-click a protein sequence** | **Toolbox** | **Protein Analyses ( )** | **Reverse translate ( )**

This opens the dialog displayed in figure 14.8:

If a sequence was selected before choosing the Toolbox action, the sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the **Project Tree**. You can translate several protein sequences at a time.

Click **Next** to adjust the parameters for the translation.

14.3.1 Reverse translation parameters

Figure 14.9 shows the choices for making the translation.

- **Most frequently used codon.** On the basis of the selected translation table, this parameter/option will assign the codon that occurs most often. When choosing this option, the results of performing several reverse translations will always be the same, contrary to the

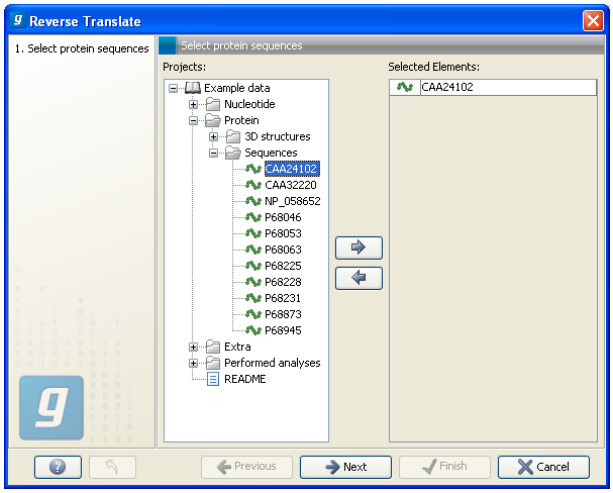


Figure 14.8: *Choosing a protein sequence for reverse translation.*

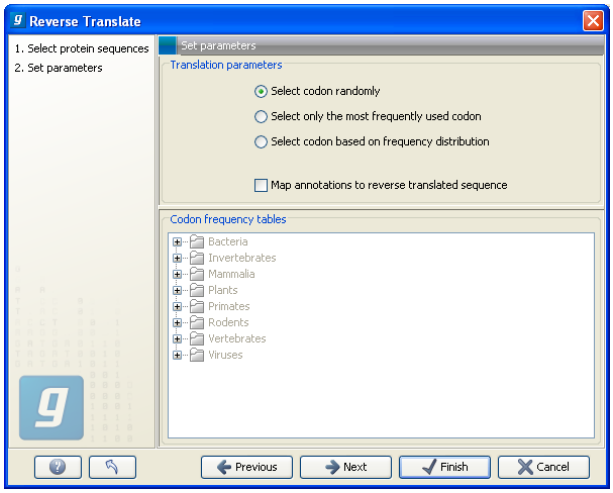


Figure 14.9: *Choosing parameters for the reverse translation.*

following two options.

- **Uniform distribution.** This parameter/option will randomly back-translate an amino acid to a codon without using the translation tables. Every time you perform the analysis you will get a different result.

- **Distribution according to frequency.** This option is a mix of the other two options. The selected translation table is used to attach weights to each codon based on its frequency. The codons are assigned randomly with a probability given by the weights. A more frequent codon has a higher probability of being selected. Every time you perform the analysis, you will get a different result. This option yields a result that is closer to the translation behavior of the organism (assuming you chose an appropriate codon frequency table).

- **Map annotations to reverse translated sequence.** If this checkbox is checked, then all annotations on the protein sequence will be mapped to the resulting DNA sequence. In the tooltip on the transferred annotations, there is a note saying that the annotation derives from the original sequence.

The **Codon Frequency Table** is used to determine the frequencies of the codons. Select a frequency table from the list that fits the organism you are working with. A translation table of an organism is created on the basis of counting all the codons in the coding sequences. Every codon in a **Codon Frequency Table** has its own count, frequency (per thousand) and fraction which are calculated in accordance with the occurrences of the codon in the organism.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. The newly created nucleotide sequence is shown, and if the analysis was performed on several protein sequences, there will be a corresponding number of views of nucleotide sequences. The new sequence is not saved automatically. To save the sequence, drag it into the **Navigation Area** or press Ctrl + S (⌘ + S on Mac) to show the save dialog.

14.3.2 Bioinformatics explained: Reverse translation

In all living cells containing hereditary material such as DNA, a transcription to mRNA and subsequent a translation to proteins occur. This is of course simplified but is in general what is happening in order to have a steady production of proteins needed for the survival of the cell. In bioinformatics analysis of proteins it is sometimes useful to know the ancestral DNA sequence in order to find the genomic localization of the gene. Thus, the translation of proteins back to DNA/RNA is of particular interest, and is called reverse translation or back-translation.

The Genetic Code

In 1968 the Nobel Prize in Medicine was awarded to Robert W. Holley, Har Gobind Khorana and Marshall W. Nirenberg for their interpretation of the Genetic Code (http://nobelprize.org/medicine/laureates/1968/. The Genetic Code represents translations of all 64 different codons into 20 different amino acids. Therefore it is no problem to translate a DNA/RNA sequence into a specific protein. But due to the degeneracy of the genetic code, several codons may code for only one specific amino acid. This can be seen in figure 14.10. After the discovery of the genetic code it has been concluded that different organism (and organelles) have genetic codes which are different from the "standard genetic code". Moreover, the amino acid alphabet is no longer limited to 20 amino acids. The 21'st amino acid, selenocysteine, is encoded by an 'UGA' codon which is normally a stop codon. The discrimination of a selenocysteine over a stop codon is carried out by the translation machinery. Selenocysteines are very rare amino acids.

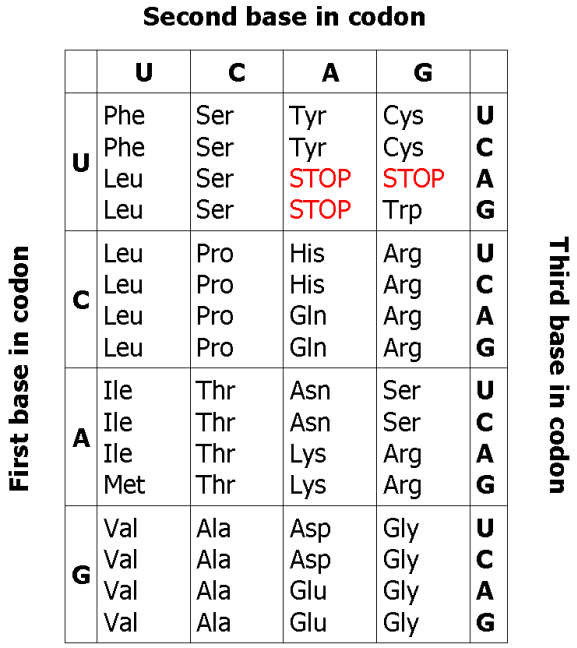
The figure 14.10 and 14.11 represents the Standard Code which is the default translation table.

```
  AAs  = FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSS**VVVVAAAADDEEGGGG
Starts = --------------------------------MMMM---------------M------------
Base1  = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2  = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3  = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

Figure 14.10: *The Standard Code for translation.*

Challenge of reverse translation

A particular protein follows from the translation of a DNA sequence whereas the reverse translation need not have a specific solution according to the Genetic Code. The Genetic Code is degenerate which means that a particular amino acid can be translated into more than one codon. Hence there are ambiguities of the reverse translation.



First base in codon	Second base in codon: U	Second base in codon: C	Second base in codon: A	Second base in codon: G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Figure 14.11: *The standard genetic code showing amino acids for all 64 possible codons.*

Solving the ambiguities of reverse translation

In order to solve these ambiguities of reverse translation you can define how to prioritize the codon selection, e.g:

- Choose a codon randomly.
- Select the most frequent codon in a given organism.
- Randomize a codon, but with respect to its frequency in the organism.

As an example we want to translate an alanine to the corresponding codon. Four different codons can be used for this reverse translation; GCU, GCC, GCA or GCG. By picking either one by random choice we will get an alanine.

The most frequent codon, coding for an alanine in *E. coli* is GCG, encoding 33.7% of all alanines. Then comes GCC (25.5%), GCA (20.3%) and finally GCU (15.3%). The data are retrieved from the Codon usage database, see below. Always picking the most frequent codon does not necessarily give the best answer.

By selecting codons from a distribution of calculated codon frequencies, the DNA sequence obtained after the reverse translation, holds the correct (or nearly correct) codon distribution. It should be kept in mind that the obtained DNA sequence is not necessarily identical to the original one encoding the protein in the first place, due to the degeneracy of the genetic code.

In order to obtain the best possible result of the reverse translation, one should use the codon frequency table from the correct organism or a closely related species. The codon usage of the mitochondrial chromosome are often different from the native chromosome(s), thus mitochondrial codon frequency tables should only be used when working specifically with mitochondria.

Other useful resources

The Genetic Code at NCBI:
http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c

Codon usage database:
http://www.kazusa.or.jp/codon/

Wikipedia on the genetic code
http://en.wikipedia.org/wiki/Genetic_code

Creative Commons License
All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and "CLC bio" has to be clearly labelled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.

See http://creativecommons.org/licenses/by-nc-nd/2.5/ for more about how you may use the contents.

Chapter 15

Primers

Contents

15.1 Primer design - an introduction . . . 185
15.1.1 General concept . . . 185
15.2 Setting parameters for primers and probes . . . 187
15.2.1 Primer Parameters . . . 187
15.3 Graphical display of primer information . . . 189
15.3.1 Compact information mode . . . 189
15.3.2 Detailed information mode . . . 190
15.4 Output from primer design . . . 191
15.4.1 Saving primers . . . 191
15.4.2 Saving PCR fragments . . . 191
15.4.3 Adding primer binding annotation . . . 192
15.5 Standard PCR . . . 192
15.5.1 User input . . . 192
15.5.2 Standard PCR output table . . . 194
15.6 Nested PCR . . . 195
15.6.1 Nested PCR output table . . . 197
15.7 TaqMan . . . 197
15.7.1 TaqMan output table . . . 199
15.8 Sequencing primers . . . 199
15.8.1 Sequencing primers output table . . . 200
15.9 Alignment-based primer and probe design . . . 200
15.9.1 Specific options for alignment-based primer and probe design . . . 200
15.9.2 Alignment based design of PCR primers . . . 201
15.9.3 Alignment-based TaqMan probe design . . . 203
15.10 Analyze primer properties . . . 204
15.11 Match primer with sequence . . . 205
15.11.1 Search for primer binding sites parameters . . . 206
15.12 Order primers . . . 207

CLC Gene Workbench offers graphically and algorithmically advanced design of primers and probes for various purposes. This chapter begins with a brief introduction to the general concepts of the primer designing process. Then follows instructions on how to adjust parameters for primers, how to inspect and interpret primer properties graphically and how to interpret, save and analyze the output of the primer design analysis. After a description of the different reaction types for which primers can be designed, the chapter closes with sections on how to match primers with other sequences and how to create a primer order.

15.1 Primer design - an introduction

Primer design can be accessed in two ways:

select sequence | **Toolbox in the Menu Bar** | **Primers and Probes (** **)** | **Design Primers(** **)** | **OK**

or **right-click sequence** | **Show** | **Primer(** **)**

In the primer view (see figure 15.1), the basic options for viewing the template sequence are the same as for the standard sequence viewer. See section 11.1 for an explanation of these options. **Notice!** This means that features such as e.g. known SNP's or exons can be displayed on the template sequence to guide the choice of primer regions. Also, traces in sequencing reads can be shown along with the structure to guide e.g. the re-sequencing of poorly resolved regions.

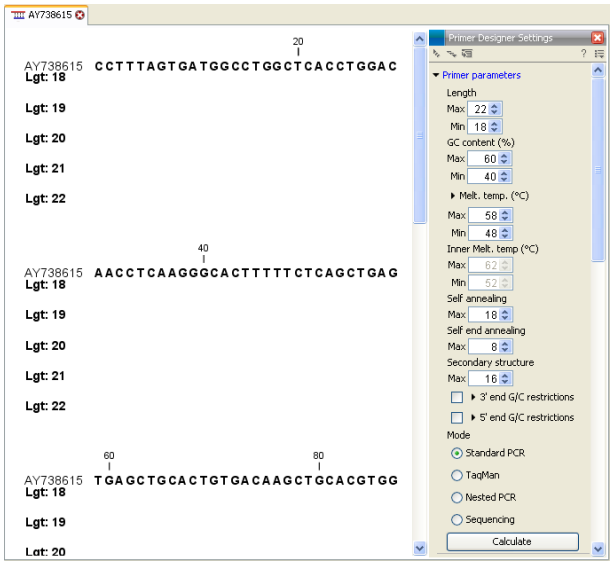


Figure 15.1: *The initial view of the sequence used for primer design.*

15.1.1 General concept

The concept of the primer view is that the user first chooses the desired reaction type for the session in the Primer Parameters preference group, e.g. *Standard PCR*. Reflecting the choice of reaction type, it is now possibly to select one or more regions on the sequence and to use the right-click mouse menu to designate these as primer or probe regions (see figure 15.2).

When a region is chosen, graphical information about the properties of all possible primers in this region will appear in lines beneath it. By default, information is showed using a compact

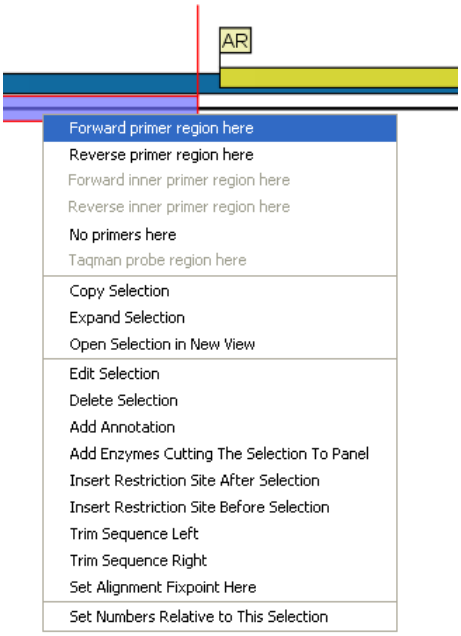


Figure 15.2: *Right-click menu allowing you to specify regions for the primer design*

mode but the user can change to a more detailed mode in the Primer information preference group.

The number of information lines reflects the chosen length interval for primers and probes. In the compact information mode one line is shown for every possible primer-length and each of these lines contain information regarding all possible primers of the given length. At each potential primer starting position a circular information point is shown which indicates whether the primer fulfills the requirements set in the primer parameters preference group. A green circle indicates a primer which fulfils all criteria and a red circle indicates a primer which fails to meet one or more of the set criteria. For more detailed information, place the mouse cursor over the circle representing the primer of interest. A tool-tip will then appear on screen, displaying detailed information about the primer in relation to the set criteria. To locate the primer on the sequence, simply left-click the circle using the mouse.

The various primer parameters can now be varied to explore their effect and the view area will dynamically update to reflect this allowing for a high degree of interactivity in the primer design process.

After having explored the potential primers the user may have found a satisfactory primer and choose to export this directly from the view area using a mouse right-click on the primers information point. This does not allow for any design information to enter concerning the properties of primer/probe pairs or sets e.g. primer pair annealing and T_m difference between primers. If the latter is desired the user can use the calculate button at the bottom of the Primer parameter preference group. This will activate a dialog, the contents of which depends on the chosen mode. Here, the user can set primer-pair specific setting such as allowed or desired T_m difference and view the single-primer parameters which were chosen in the Primer parameters preference group.

Upon pressing finish, an algorithm will generate all possible primer sets and rank these based on their characteristics and the chosen parameters. A list will appear displaying the 100 most

high scoring sets and information pertaining to these. The search result can be saved to the navigator. From the result table, suggested primers or primer/probe sets can be explored since clicking an entry in the table will highlight the associated primers and probes on the sequence. It is also possible to save individual primers or sets from the table through the mouse right-click menu. For a given primer pair, the amplified PCR fragment can also be opened or saved using the mouse right-click menu.

15.2 Setting parameters for primers and probes

The primer-specific view options and settings are found in the **Primer parameters** preference group in the **Side Panel** to the right of the view (see figure 15.3).

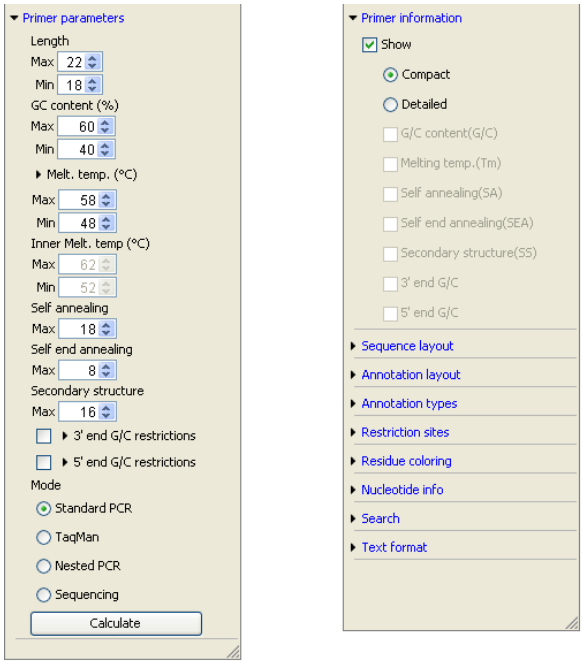


Figure 15.3: *The two groups of primer parameters (in the program, the Primer information group is listed below the other group).*

15.2.1 Primer Parameters

In this preference group a number of criteria can be set, which the selected primers must meet. All the criteria concern *single primers*, as primer pairs are not generated until the calculate button is pressed. Parameters regarding primer and probe sets are described in detail for each reaction mode (see below).

- **Length.** Determines the length interval within which primers can be designed by setting a maximum and a minimum length. The upper and lower lengths allowed by the program are 50 and 10 nucleotides respectively.
- **GC content.** Determines the interval of CG content (% C and G nucleotides in the primer) within which primers must lie by setting a maximum and a minimum GC content.

- **Melting temperature.** Determines the temperature interval within which primers must lie. When the *Nested PCR* or *TaqMan* reaction type is chosen, the first pair of melting temperature interval settings relate to the outer primer pair i.e. not the probe. The melting temperature group can also be unfolded to show parameters regarding the reaction mixture:
  - **Primer concentration.** Specifies the concentration of primers and probes in units of nanomoles (nM)
  - **Salt concentration.** Specifies the concentration of monovalent cations ($[NA^+]$, $[K^+]$ and equivalents) in units of millimoles (mM)

  Melting temperatures are calculated by a nearest-neighbor model which considers stacking interactions between neighboring bases in the primer-template complex. The model uses state-of-the-art thermodynamic parameters [SantaLucia, 1998] and considers the important contribution from the dangling ends that are present when a short primer anneals to a template sequence [Bommarito et al., 2000]

- **Inner melting temperature.** This option is only activated when the *Nested PCR* or *TaqMan* mode is selected. In *Nested PCR* mode, it determines the allowed melting temperature interval for the inner/nested pair of primers, and in *TaqMan* mode it determines the allowed temperature interval for the TaqMan probe.

- **Self annealing.** Determines the maximum self annealing value of all primers and probes. This determines the amount of base-pairing allowed between two copies of the same molecule. The self annealing score is measured in number of hydrogen bonds between two copies of primer molecules, with A-T base pairs contributing 2 hydrogen bonds and G-C base pairs contributing 3 hydrogen bonds.

- **Self end annealing.** Determines the maximum self end annealing value of all primers and probes. This determines the amount of consecutive base pairs allowed between the ends of two copies of the same molecule. This score is also calculated in units of hydrogen bonds between two primer copies of identical primer molecules.

- **Secondary structure.** Determines the maximum score of the optimal secondary DNA structure found for a primer or probe. Secondary structures are scored by the number of hydrogen bonds in the structure, and 2 extra hydrogen bonds are added for each stacking base-pair in the structure.

- **3' end G/C restrictions.** When this checkbox is selected it is possible to specify restrictions concerning the number of G and C molecules in the 3' end of primers and probes. A low G/C content of the primer/probe 3' end increases the specificity of the reaction. A high G/C content facilitates a tight binding of the oligo to the template but also increases the possibility of mispriming. Unfolding the preference groups yields the following options:
  - **End length.** The number of consecutive terminal nucleotides for which to consider the C/G content
  - **Max no. of G/C.** The maximum number of G and C nucleotides allowed within the specified length interval
  - **Min no. of G/C.** The minimum number of G and C nucleotides required within the specified length interval

- **5' end G/C restrictions.** When this checkbox is selected it is possible to specify restrictions concerning the number of G and C molecules in the 5' end of primers and probes. A high G/C content facilitates a tight binding of the oligo to the template but also increases the possibility of mis-priming. Unfolding the preference groups yields the same options as described above for the 3' end.
- **Mode.** Specifies the reaction type for which primers are designed
  - **Standard PCR.** Used when the objective is to design primers, or primer pairs, for PCR amplification of a single DNA fragment.
  - **Nested PCR.** Used when the objective is to design two primer pairs for nested PCR amplification of a single DNA fragment.
  - **Sequencing.** Used when the objective is to design primers for DNA sequencing.
  - **TaqMan.** Used when the objective is to design a primer pair and a probe for TaqMan quantitative PCR.

  Each mode is described further below.
- **Calculate.** Pushing this button will activate the algorithm for designing primers

15.3 Graphical display of primer information

The primer information settings are found in the **Primer information** preference group in the **Side Panel** to the right of the view (see figure 15.3).

There are two different ways to display the information relating to a single primer, the detailed and the compact view. Both are shown below the primer regions selected on the sequence.

15.3.1 Compact information mode

This mode offers a condensed overview of all the primers that are available in the selected region. When a region is chosen primer information will appear in lines beneath it (see figure 15.4).

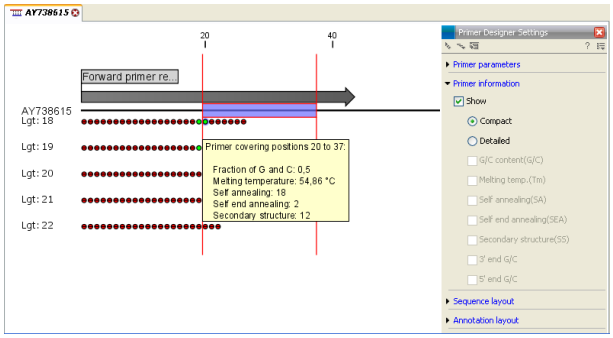


Figure 15.4: *Compact information mode*

The number of information lines reflects the chosen length interval for primers and probes. One line is shown for every possible primer-length, if the length interval is widened more lines will appear. At each potential primer starting position a circle is shown which indicates whether the primer fulfills the requirements set in the primer parameters preference group. A green primer indicates a primer which fulfils all criteria and a red primer indicates a primer which fails to meet one or more of the set criteria. For more detailed information, place the mouse cursor over the

circle representing the primer of interest. A tool-tip will then appear on screen displaying detailed information about the primer in relation to the set criteria. To locate the primer on the sequence, simply left-click the circle using the mouse.

The various primer parameters can now be varied to explore their effect and the view area will dynamically update to reflect this. If e.g. the allowed melting temperature interval is widened more green circles will appear indicating that more primers now fulfill the set requirements and if e.g. a requirement for 3' G/C content is selected, rec circles will appear at the starting points of the primers which fail to meet this requirement.

15.3.2 Detailed information mode

In this mode a very detailed account is given of the properties of all the available primers. When a region is chosen primer information will appear in groups of lines beneath it (see figure 15.5).

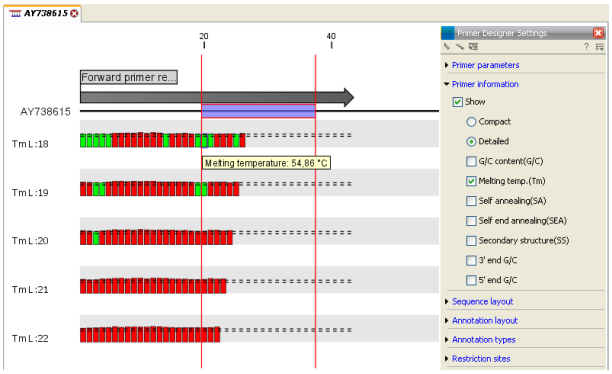


Figure 15.5: *Detailed information mode*

The number of information-line-groups reflects the chosen length interval for primers and probes. One group is shown for every possible primer length. Within each group, a line is shown for every primer property that is selected from the checkboxes in the primer information preference group. Primer properties are shown at each potential primer starting position and are of two types:

Properties with numerical values are represented by bar plots. A green bar represents the starting point of a primer that meets the set requirement and a red bar represents the starting point of a primer that fails to meet the set requirement:

- G/C content
- Melting temperature
- Self annealing score
- Self end annealing score
- Secondary structure score

Properties with Yes - No values. If a primer meets the set requirement a green circle will be shown at its starting position and if it fails to meet the requirement a red dot is shown at its starting position:

- C/G at 3' end

- C/G at 5’ end

Common to both sorts of properties is that mouse clicking an information point (filled circle or bar) will cause the region covered by the associated primer to be selected on the sequence.

15.4 Output from primer design

The output generated by the primer design algorithm is a table of proposed primers or primer pairs with the accompanying information (see figure 15.6).

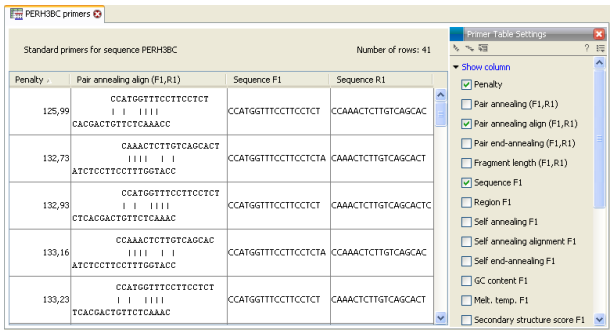


Figure 15.6: *Proposed primers*

In the preference panel of the table, it is possible to customize which columns are shown in the table. See the sections below on the different reaction types for a description of the available information.

The columns in the output table can be sorted by the present information. For example the user can choose to sort the available primers by their penalty value (default) or by their self annealing score, simply by right-clicking the column header.

The output table interacts with the accompanying primer editor such that when a proposed combination of primers and probes is selected in the table the primers and probes in this solution are highlighted on the sequence.

15.4.1 Saving primers

Primer solutions in a table row can be saved by selecting the row and using the right-click mouse menu. This opens a dialogue that allows the user to save the primers to the desired location. Primers and probes are saved as DNA sequences in the program. This means that all available DNA analyzes can be performed on the saved primers, including BLAST. Furthermore, the primers can be edited using the standard sequence viewer to introduce e.g. mutations and restriction sites.

15.4.2 Saving PCR fragments

The PCR fragment generated from the primer pair in a given table row can also be saved by selecting the row and using the right-click mouse menu. This opens a dialogue that allows the user to save the fragment to the desired location. The fragment is saved as a DNA sequence and the position of the primers is added as annotation on the sequence. The fragment can then be used for further analysis and included in e.g. an in-silico cloning experiment using the cloning editor.

15.4.3 Adding primer binding annotation

You can add an annotation to the template sequence specifying the binding site of the primer: Right-click the primer in the table and select **Mark primer annotation on sequence**.

15.5 Standard PCR

This mode is used to design primers for a PCR amplification of a single DNA fragment.

15.5.1 User input

In this mode the user must define either a *Forward primer region*, a *Reverse primer region*, or both. These are defined using the mouse right-click menu. If areas are known where primers must not bind (e.g. repeat rich areas), one or more *No primers here* regions can be defined.

If two regions are defined, it is required that the *Forward primer region*, is located upstream of the *Reverse primer region*.

After exploring the available primers (see 15.3) and setting the desired parameter values in the Primer Parameters preference group, the calculate button will activate the primer design algorithm.

If only a single region is defined only single primers will be suggested by the algorithm.

After pressing the calculate button a dialogue will appear (see figure 15.7).

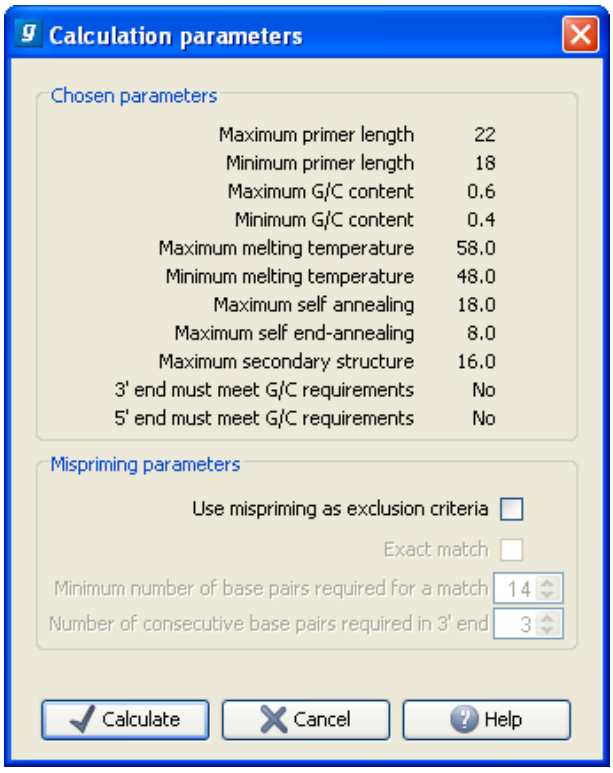

Calculation parameters
Chosen parameters
Maximum primer length 22
Minimum primer length 18
Maximum G/C content 0.6
Minimum G/C content 0.4
Maximum melting temperature 58.0
Minimum melting temperature 48.0
Maximum self annealing 18.0
Maximum self end-annealing 8.0
Maximum secondary structure 16.0
3' end must meet G/C requirements No
5' end must meet G/C requirements No
Mispriming parameters
Use mispriming as exclusion criteria
Exact match
Minimum number of base pairs required for a match 14
Number of consecutive base pairs required in 3' end 3
Calculate
Cancel
Help

Figure 15.7: *Calculation dialog for PCR primers*

The top part of this dialog shows the parameter settings chosen in the Primer parameters preference group which will be used by the design algorithm.

The lower part contains a menu where the user can choose to include mispriming as a criteria in the design process. If this option is selected the algorithm will search for competing binding sites of the primer within the sequence.

The adjustable parameters for the search are:

- **Exact match**. Choose only to consider exact matches of the primer, i.e. all positions must base pair with the template for mispriming to occur.
- **Minimum number of base pairs required for a match**. How many nucleotides of the primer that must base pair to the sequence in order to cause mispriming.
- **Number of consecutive base pairs required in 3' end**. How many consecutive 3' end base pairs in the primer that MUST be present to for mispriming to occur. This option is included since 3' terminal base pairs are known to be essential for priming to occur.

Notice! that including a search for potential mispriming sites will prolong the search time substantially if long sequences are used as template and if the minimum number of base pairs required for a match is low. If the region to be amplified is part of a very long molecule and mispriming is a concern, consider extracting part of the sequence prior to designing primers.

If both a forward and a reverse region are defined primer pairs will be suggested by the algorithm.

After pressing the calculate button a dialogue will appear (see figure 15.8).

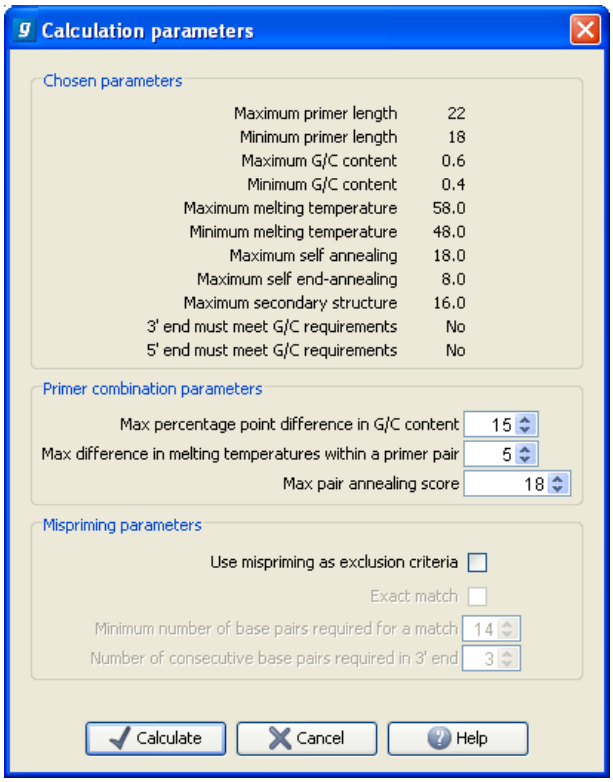

Calculation parameters
Chosen parameters
Maximum primer length 22
Minimum primer length 18
Maximum G/C content 0.6
Minimum G/C content 0.4
Maximum melting temperature 58.0
Minimum melting temperature 48.0
Maximum self annealing 18.0
Maximum self end-annealing 8.0
Maximum secondary structure 16.0
3' end must meet G/C requirements No
5' end must meet G/C requirements No
Primer combination parameters
Max percentage point difference in G/C content 15
Max difference in melting temperatures within a primer pair 5
Max pair annealing score 18
Mispriming parameters
Use mispriming as exclusion criteria
Exact match
Minimum number of base pairs required for a match 14
Number of consecutive base pairs required in 3' end 3
Calculate
Cancel
Help

Figure 15.8: *Calculation dialog for PCR primers*

Again, the top part of this dialog shows the parameter settings chosen in the Primer parameters preference group which will be used by the design algorithm. The lower part again contains a

menu where the user can choose to include mispriming of both primers as a criteria in the design process (see above). The central part of the dialogue contains parameters pertaining to primer pairs. Here three parameters can be set:

- Maximum percentage point difference in G/C content - if this is set at e.g. 5 points a pair of primers with 45% and 49% G/C nucleotides, respectively, will be allowed, whereas a pair of primers with 45% and 51% G/C nucleotides, respectively will not be included.
- Maximal difference in melting temperature of primers in a pair - the number of degrees Celsius that primers in a pair are all allowed to differ.
- Maximum pair annealing score - the maximum number of hydrogen bonds allowed between the forward and the reverse primer in a primer pair.

15.5.2 Standard PCR output table

If only a single region is selected the following columns of information are available:

- Sequence - the primer's sequence.
- Penalty - measures how much the properties of the primer (or primer pair) deviates from the optimal solution in terms of the chosen parameters. The lower the penalty, the better the solution.
- Region - the interval of the template sequence covered by the primer
- Self annealing - the maximum self annealing score of the primer in units of hydrogen bonds
- Self annealing alignment - a visualization of the highest maximum scoring self annealing alignment
- Self end annealing - the maximum score of consecutive end base-pairings allowed between the ends of two copies of the same molecule in units of hydrogen bonds
- GC content - the fraction of G and C nucleotides in the primer
- Melting temperature of the primer-template complex
- Secondary structure score - the score of the optimal secondary DNA structure found for the primer. Secondary structures are scored by adding the number of hydrogen bonds in the structure, and 2 extra hydrogen bonds are added for each stacking base-pair in the structure
- Secondary structure - a visualization of the optimal DNA structure found for the primer

If both a forward and a reverse region are selected a table of primer pairs is shown, where the above columns (excluding the penalty) are represented twice, once for the forward primer (designated by the letter F) and once for the reverse primer (designated by the letter R).

Before these, and following the penalty of the primer pair, are the following columns pertaining to primer pair-information available:

- Pair annealing - the number of hydrogen bonds found in the optimal alignment of the forward and the reverse primer in a primer pair
- Pair annealing alignment - a visualization of the optimal alignment of the forward and the reverse primer in a primer pair.
- Pair end annealing - the maximum score of consecutive end base-pairings found between the ends of the two primers in the primer pair, in units of hydrogen bonds
- Fragment length - the length (number of nucleotides) of the PCR fragment generated by the primer pair

15.6 Nested PCR

Nested PCR is a modification of Standard PCR, aimed at reducing product contamination due to the amplification of unintended primer binding sites (mispriming). If the intended fragment can not be amplified without interference from competing binding sites, the idea is to seek out a larger outer fragment which can be unambiguously amplified and which contains the smaller intended fragment. Having amplified the outer fragment to large numbers, the PCR amplification of the inner fragment can proceed and will yield amplification of this with minimal contamination.

Primer design for nested PCR thus involves designing two primer pairs, one for the outer fragment and one for the inner fragment.

In *Nested PCR* mode the user must thus define four regions a *Forward primer region* (the outer forward primer), a *Reverse primer region* (the outer reverse primer), a *Forward inner primer region*, and a *Reverse inner primer region*. These are defined using the mouse right-click menu. If areas are known where primers must not bind (e.g. repeat rich areas), one or more *No primers here* regions can be defined.

It is required that the *Forward primer region*, is located upstream of the *Forward inner primer region*, that the *Forward inner primer region*, is located upstream of the *Reverse inner primer region*, and that the *Reverse inner primer region*, is located upstream of the *Reverse primer region*.

In *Nested PCR* mode the *Inner melting temperature* menu in the Primer parameters panel is activated, allowing the user to set a separate melting temperature interval for the inner and outer primer pairs.

After exploring the available primers (see section 15.3) and setting the desired parameter values in the Primer parameters preference group, the calculate button will activate the primer design algorithm.

After pressing the calculate button a dialogue will appear (see figure 15.9).

The top and bottom parts of this dialog are identical to the *Standard PCR* dialogue for designing primer pairs described above.

The central part of the dialogue contains parameters pertaining to primer pairs and the comparison between the outer and the inner pair. Here five options can be set:

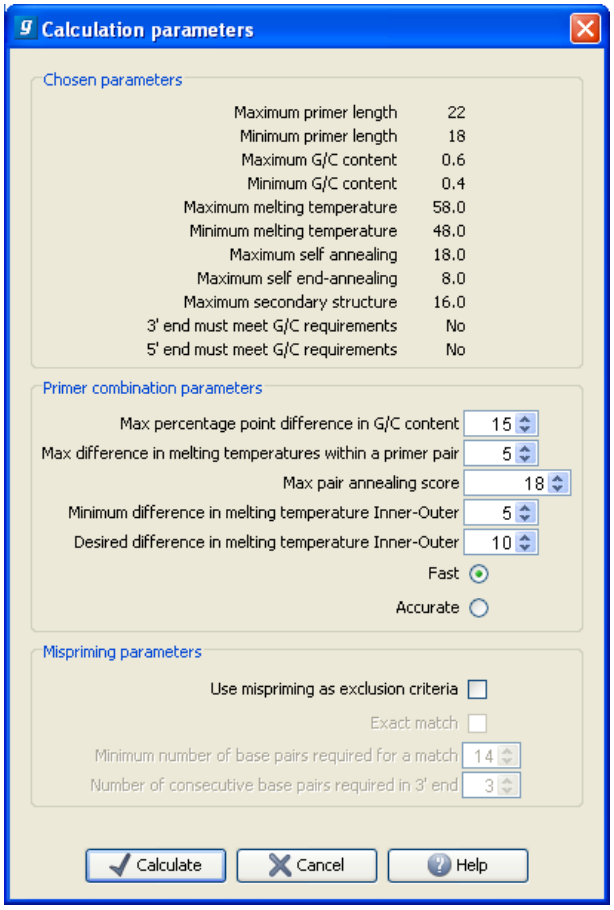

Calculation parameters
Chosen parameters
Maximum primer length 22
Minimum primer length 18
Maximum G/C content 0.6
Minimum G/C content 0.4
Maximum melting temperature 58.0
Minimum melting temperature 48.0
Maximum self annealing 18.0
Maximum self end-annealing 8.0
Maximum secondary structure 16.0
3' end must meet G/C requirements No
5' end must meet G/C requirements No
Primer combination parameters
Max percentage point difference in G/C content 15
Max difference in melting temperatures within a primer pair 5
Max pair annealing score 18
Minimum difference in melting temperature Inner-Outer 5
Desired difference in melting temperature Inner-Outer 10
Fast
Accurate
Mispriming parameters
Use mispriming as exclusion criteria
Exact match
Minimum number of base pairs required for a match 14
Number of consecutive base pairs required in 3' end 3
Calculate
Cancel
Help

Figure 15.9: *Calculation dialog*

- Maximum percentage point difference in G/C content (described above under Standard PCR) - this criteria is applied to both primer pairs independently.
- Maximal difference in melting temperature of primers in a pair - the number of degrees Celsius that primers in a pair are all allowed to differ. This criteria is applied to both primer pairs independently.
- Maximum pair annealing score - the maximum number of hydrogen bonds allowed between the forward and the reverse primer in a primer pair. This criteria is applied to all possible combinations of primers.
- Minimum difference in the melting temperature of primers in the inner and outer primer pair - all comparisons between the melting temperature of primers from the two pairs must be at least this different, otherwise the primer set is excluded. This option is applied to ensure that the inner and outer PCR reactions can be initiated at different annealing temperatures.
- Desired temperature difference in melting temperature between outer and inner primer pair - the scoring function discounts primer sets which deviate greatly from this value. Regarding this , and the minimum difference option mentioned above, please note that to ensure flexibility there is no directionality indicated when setting parameters for melting temperature differences between inner and outer primer pair, i.e. it is not specified whether the inner pair should have a lower or higher T_m. Instead this is determined by the allowed temperature intervals for inner and outer primers that are set in the primer parameters

preference group in the side panel. If a higher T_m of inner primers is desired, choose a T_m interval for inner primers which has higher values than the interval for outer primers.

- Two radio buttons allowing the user to choose between a fast and an accurate algorithm for primer prediction.

15.6.1 Nested PCR output table

In nested PCR there are four primers in a solution, forward outer primer (FO), forward inner primer (FI), reverse inner primer (RI) and a reverse outer primer (RO).

The output table can show primer-pair combination parameters for all four combinations of primers and single primer parameters for all four primers in a solution (see section on Standard PCR for an explanation of the available primer-pair and single primer information).

The fragment length in this mode refers to the length of the PCR fragment generated by the inner primer pair, and this is also the PCR fragment which can be exported.

15.7 TaqMan

CLC Gene Workbench allows the user to design primers and probes for TaqMan PCR applications.

TaqMan probes are oligonucleotides that contain a fluorescent reporter dye at the 5' end and a quenching dye at the 3' end. Fluorescent molecules become excited when they are irradiated and usually emit light. However, in a TaqMan probe the energy form the fluorescent dye is transferred to the quencher dye by fluorescence resonance energy transfer as long as the quencher and the dye are located in close proximity i.e. when the probe is intact. TaqMan probes are designed to anneal within a PCR product amplified by a standard PCR primer pair. If a TaqMan probe is bound to a product template, the replication of this will cause the Taq polymerase to encounter the probe. Upon doing so, the 5'exonuclease activity of the polymerase will cleave the probe. This cleavage separates the quencher and the dye, and as a result the reporter dye starts to emit fluorescence.

The TaqMan technology is used in Real-Time quantitative PCR. Since the accumulation of fluorescence mirrors the accumulation of PCR products it can can be monitored in real-time and used to quantify the amount of template initially present in the buffer.

The technology is also used to detect genetic variation such as SNP's. By designing a TaqMan probe which will specifically bind to one of two or more genetic variants it is possible to detect genetic variants by the presence or absence of fluorescence in the reaction. **Notice!** in *CLC Gene Workbench* it is possible to annotate sequences with SNP information from dbSNP and use this information to guide TaqMan allele-specific probe design.

A specific requirement of TaqMan probes is that a G nucleotide can not be present at the 5' end since this will quench the fluorescence of the reporter dye. It is recommended that the melting temperature of the TaqMan probe is about 10 degrees celsius higher than that of the primer pair.

Primer design for TaqMan technology involves designing a primer pair and a TaqMan probe.

In *TaqMan* mode the user must thus define three regions a *Forward primer region*, a *Reverse primer region*, and a *TaqMan Probe region*. These are defined using the mouse right-click menu.

The *TaqMan Probe region* is per default oriented forward, i.e. made so that it binds to the complementary strand of the displayed single stranded template. To obtain a *TaqMan Probe region* which is oriented in the reverse direction, chose the double stranded sequence layout and select the complementary strand before defining the *TaqMan Probe region*. If areas are known where primers or probes must not bind (e.g. repeat rich areas), one or more *No primers here* regions can be defined.

It is required that the *Forward primer region*, is located upstream of the *TaqMan Probe region*, and that the *TaqMan Probe region*, is located upstream of the *Reverse primer region*.

In *TaqMan* mode the *Inner melting temperature* menu in the primer parameters panel is activated allowing the user to set a separate melting temperature interval for the TaqMan probe.

After exploring the available primers (see section 15.3) and setting the desired parameter values in the Primer Parameters preference group, the calculate button will activate the primer design algorithm.

After pressing the calculate button a dialogue will appear (see 15.10) which is identical to the *Nested PCR* dialogue described above (see section 15.6).

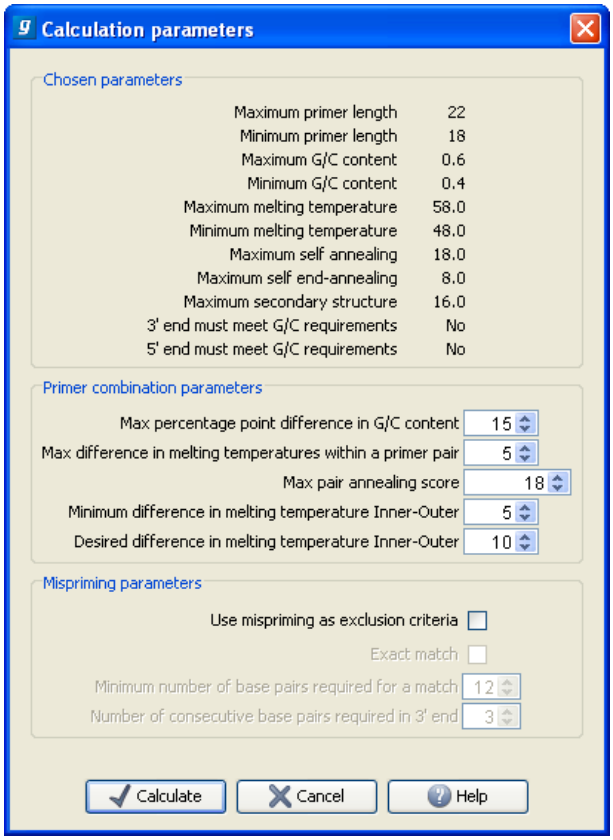

Calculation parameters
Chosen parameters
Maximum primer length 22
Minimum primer length 18
Maximum G/C content 0.6
Minimum G/C content 0.4
Maximum melting temperature 58.0
Minimum melting temperature 48.0
Maximum self annealing 18.0
Maximum self end-annealing 8.0
Maximum secondary structure 16.0
3' end must meet G/C requirements No
5' end must meet G/C requirements No
Primer combination parameters
Max percentage point difference in G/C content 15
Max difference in melting temperatures within a primer pair 5
Max pair annealing score 18
Minimum difference in melting temperature Inner-Outer 5
Desired difference in melting temperature Inner-Outer 10
Mispriming parameters
Use mispriming as exclusion criteria
Exact match
Minimum number of base pairs required for a match 12
Number of consecutive base pairs required in 3' end 3
Calculate
Cancel
Help

Figure 15.10: *Calculation dialog*

In this dialogue the options to set a minimum and a desired melting temperature difference between outer and inner refers to primer pair and probe respectively.

15.7.1 TaqMan output table

In TaqMan mode there are two primers and a probe in a given solution, forward primer (F), reverse primer (R) and a TaqMan probe (TP).

The output table can show primer/probe-pair combination parameters for all three combinations of primers and single primer parameters for both primers and the TaqMan probe (see section on Standard PCR for an explanation of the available primer-pair and single primer information).

The fragment length in this mode refers to the length of the PCR fragment generated by the primer pair, and this is also the PCR fragment which can be exported.

15.8 Sequencing primers

This mode is used to design primers for DNA sequencing.

In this mode the user can define a number of *Forward primer regions* and *Reverse primer regions* where a sequencing primer can start. These are defined using the mouse right-click menu. If areas are known where primers must not bind (e.g. repeat rich areas), one or more *No primers here* regions can be defined.

No requirements are instated on the relative position of the regions defined.

After exploring the available primers (see section 15.3) and setting the desired parameter values in the Primer Parameters preference group, the calculate button will activate the primer design algorithm.

After pressing the calculate button a dialogue will appear (see figure 15.11).

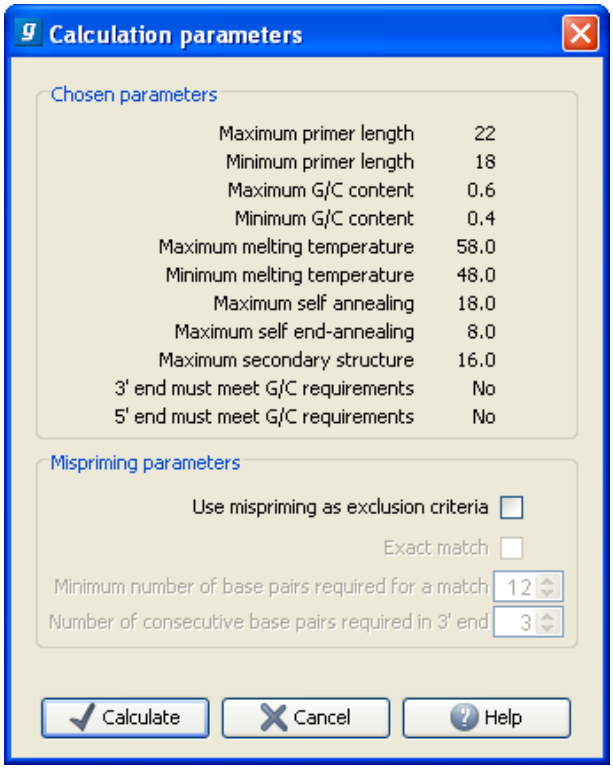

Calculation parameters
Chosen parameters
Maximum primer length 22
Minimum primer length 18
Maximum G/C content 0.6
Minimum G/C content 0.4
Maximum melting temperature 58.0
Minimum melting temperature 48.0
Maximum self annealing 18.0
Maximum self end-annealing 8.0
Maximum secondary structure 16.0
3' end must meet G/C requirements No
5' end must meet G/C requirements No
Mispriming parameters
Use mispriming as exclusion criteria
Exact match
Minimum number of base pairs required for a match 12
Number of consecutive base pairs required in 3' end 3
Calculate
Cancel
Help

Figure 15.11: *Calculation dialog for sequencing primers*

Since design of sequencing primers does not require the consideration of interactions between primer pairs, this dialogue is identical to the dialogue shown in *Standard PCR* mode when only a single primer region is chosen. See the section 15.5 for a description.

15.8.1 Sequencing primers output table

In this mode primers are predicted independently for each region, but the optimal solutions are all presented in one table. The solutions are numbered consecutively according to their position on the sequence such that the forward primer region closest to the 5' end of the molecule is designated F1, the next one F2 etc.

For each solution, the single primer information described under Standard PCR is available in the table.

15.9 Alignment-based primer and probe design

CLC Gene Workbench 2.0 allows the user to design PCR primers and TaqMan probes based on an alignment of multiple sequences.

The primer designer for alignments can be accessed in two ways:

select alignment | **Toolbox** | **Primers and Probes (** **)** | **Design Primers(** **)** | **OK**

or **right-click alignment** | **Show** | **Primer Design(** **)**

In the alignment primer view (see figure 15.12), the basic options for viewing the template alignment are the same as for the standard view of alignments. See section 18 for an explanation of these options. **Notice!** This means that features such as e.g. known SNP's or exons can be displayed on the template sequence to guide the choice of primer regions. Since the definition of groups of sequences is essential to the primer design the selection boxes of the standard view are shown as default in the alignment primer view.

15.9.1 Specific options for alignment-based primer and probe design

Compared to the primer view of a single sequence the most notable difference is that the alignment primer view has no available graphical information. Furthermore, the selection boxes found to the right of the names in the alignment play an important role in specifying the oligo design process. This is elaborated below. The Primer Parameters preference group has the same options for specifying primer requirements, but differs by the following (see figure 15.12):

- In the **Mode** submenu which specifies the reaction types the following options are found:
  - **Standard PCR.** Used when the objective is to design primers, or primer pairs, for PCR amplification of a single DNA fragment.
  - **TaqMan.** Used when the objective is to design a primer pair and a probe set for TaqMan quantitative PCR.
- The **Primer solution** submenu is used to specify requirements for the match of a PCR primer against the template sequences. These options are described further below. It contains the following options:

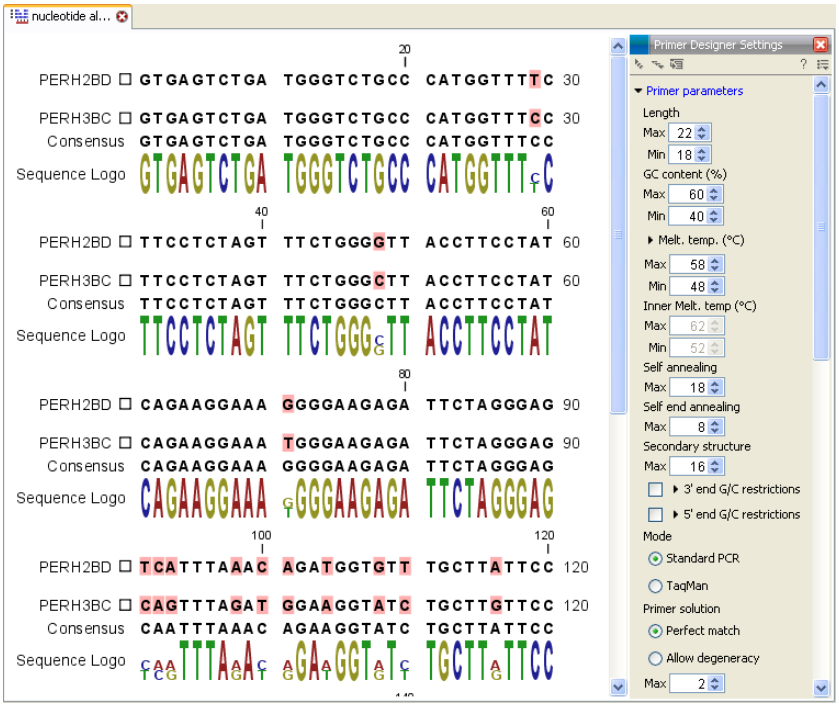


Figure 15.12: *The initial view of an alignment used for primer design.*

- **Perfect match.**
- **Allow degeneracy.**
- **Allow mismatches.**

The work flow when designing alignment based primers and probes is as follows:

- Use selection boxes to specify groups of included and excluded sequences.
- Mark either a single forward primer region, a single revers primer region or both on the sequence (and perhaps also a TaqMan region). Selections must cover all sequences in the included group.
- Adjust parameters regarding single primers in the preference panel.
- Push the Calculate button.

15.9.2 Alignment based design of PCR primers

In this mode, a single or a pair of PCR primers are designed. *CLC Gene Workbench* allows the user to design primers which will specifically amplify a group of *included* sequences but **not** amplify the remainder of the sequences, the *excluded* sequences. The selection boxes are used to indicate the status of a sequence, if the box is checked the sequence belongs to the included sequences, if not, it belongs to the excluded sequences. To design primers that are general for all primers in an alignment, simply add them all to the set of included sequences by checking all selection boxes. Specificity of priming is determined by criteria set by the user in the dialog box which is shown when the Calculate button is pressed (see below).

Different options can be chosen concerning the match of the primer to the template sequences in the included group:

- **Perfect match.** Specifies that the designed primers must have a perfect match to all relevant sequences in the alignment. When selected, primers will thus only be located in regions that are completely conserved within the sequences belonging to the included group.
- **Allow degeneracy.** Designs primers that may include ambiguity characters where heterogeneities occur in the included template sequences. The allowed fold of degeneracy is user defined and corresponds to the number of possible primer combinations formed by a degenerate primer. Thus, if a primer covers two 4-fold degenerate site and one 2-fold degenerate site the total fold of degeneracy is $4 * 4 * 2 = 32$ and the primer will, when supplied from the manufacturer, consist of a mixture of 32 different oligonucleotides. When scoring the available primers, degenerate primers are given a penalty which increases with the fold of degeneracy.
- **Allow mismatches.** Designs primers which are allowed a specified number of mismatches to the included template sequences. The melting temperature algorithm employed includes the latest thermodynamic parameters for calculating T_m when single-base mismatches occur.

When in Standard PCR mode, pushing the Calculate button will prompt the dialog shown in figure 15.13.

The top part of this dialog shows the single-primer parameter settings chosen in the Primer parameters preference group which will be used by the design algorithm.

The central part of the dialogue contains parameters pertaining to primer specificity (this is omitted if all sequences belong to the included group). Here, three parameters can be set:

- Minimum number of mismatches - the minimum number of mismatches that a primer must have against all sequences in the excluded group to ensure that it does not prime these.
- Minimum number of mismatches in 3' end - the minimum number of mismatches that a primer must have in its 3' end against all sequences in the excluded group to ensure that it does not prime these.
- Length of 3' end - the number of consecutive nucleotides to consider for mismatches in the 3' end of the primer.

The lower part of the dialog contains parameters pertaining to primer pairs (this is omitted when only designing a single primer). Here, three parameters can be set:

- Maximum percentage point difference in G/C content - if this is set at e.g. 5 points a pair of primers with 45% and 49% G/C nucleotides, respectively, will be allowed, whereas a pair of primers with 45% and 51% G/C nucleotides, respectively will not be included.
- Maximal difference in melting temperature of primers in a pair - the number of degrees Celsius that primers in a pair are all allowed to differ.
- Maximum pair annealing score - the maximum number of hydrogen bonds allowed between the forward and the reverse primer in a primer pair.

The output of the design process is a table of single primers or primer pairs as described for primer design based on single sequences. These primers are specific to the included sequences in the alignment according to the criteria defined for specificity. The only novelty in the table, is that melting temperatures are displayed with both a maximum, a minimum and an average value to reflect that degenerate primers or primers with mismatches may have heterogeneous behavior on the different templates in the group of included sequences.

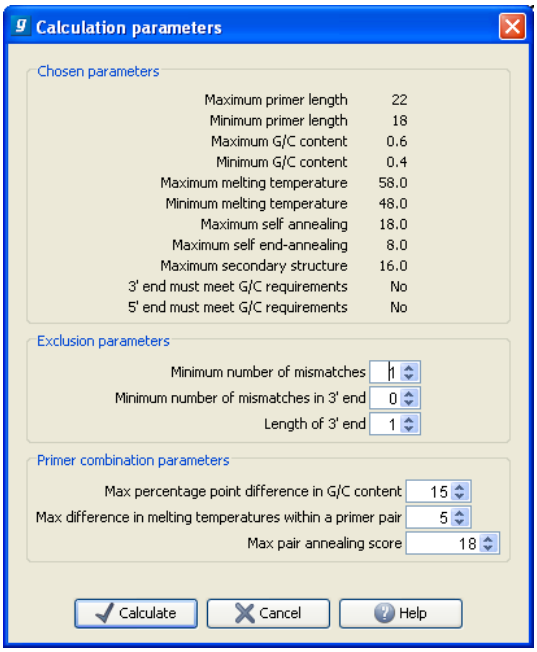

Calculation parameters
Chosen parameters
Maximum primer length 22
Minimum primer length 18
Maximum G/C content 0.6
Minimum G/C content 0.4
Maximum melting temperature 58.0
Minimum melting temperature 48.0
Maximum self annealing 18.0
Maximum self end-annealing 8.0
Maximum secondary structure 16.0
3' end must meet G/C requirements No
5' end must meet G/C requirements No
Exclusion parameters
Minimum number of mismatches 1
Minimum number of mismatches in 3' end 0
Length of 3' end 1
Primer combination parameters
Max percentage point difference in G/C content 15
Max difference in melting temperatures within a primer pair 5
Max pair annealing score 18
Calculate
Cancel
Help

Figure 15.13: *Calculation dialog shown when designing alignment based PCR primers.*

15.9.3 Alignment-based TaqMan probe design

CLC Gene Workbench allows the user to design solutions for TaqMan quantitative PCR which consist of four oligos: a general primer pair which will amplify all sequences in the alignment, a specific TaqMan probe which will match the group of *included* sequences but **not** match the *excluded* sequences and a specific TaqMan probe which will match the group of *excluded* sequences but **not** match the *included* sequences. As above, the selection boxes are used to indicate the status of a sequence, if the box is checked the sequence belongs to the included sequences, if not, it belongs to the excluded sequences. We use the terms included and excluded here to be consistent with the section above although a probe solution is presented for both groups. In TaqMan mode, primers are not allowed degeneracy or mismatches to any template sequence in the alignment, variation is only allowed/required in the TaqMan probes.

Pushing the Calculate button will cause the dialog shown in figure 15.14 to appear.

The top part of this dialog is identical to the *Standard PCR* dialog for designing primer pairs described above.

The central part of the dialog contains parameters to define the specificity of TaqMan probes. Two parameters can be set:

- Minimum number of mismatches - the minimum total number of mismatches that must exist between a specific TaqMan probe and all sequences which belong to the group not recognized by the probe.

- Minimum number of mismatches in central part - the minimum number of mismatches in the central part of the oligo that must exist between a specific TaqMan probe and all sequences which belong to the group not recognized by the probe.

The lower part of the dialogue contains parameters pertaining to primer pairs and the comparison between the outer oligos(primers) and the inner oligos (TaqMan probes). Here, five options can be set:

- Maximum percentage point difference in G/C content (described above under Standard PCR).
- Maximal difference in melting temperature of primers in a pair - the number of degrees Celsius that primers in the primer pair are all allowed to differ.
- Maximum pair annealing score - the maximum number of hydrogen bonds allowed between the forward and the reverse primer in an oligo pair. This criteria is applied to all possible combinations of primers and probes.
- Minimum difference in the melting temperature of primer (outer) and TaqMan probe (inner) oligos - all comparisons between the melting temperature of primers and probes must be at least this different, otherwise the solution set is excluded.
- Desired temperature difference in melting temperature between outer (primers) and inner (TaqMan) oligos - the scoring function discounts solution sets which deviate greatly from this value. Regarding this, and the minimum difference option mentioned above, please note that to ensure flexibility there is no directionality indicated when setting parameters for melting temperature differences between probes and primers, i.e. it is not specified whether the probes should have a lower or higher T_m. Instead this is determined by the allowed temperature intervals for inner and outer oligos that are set in the primer parameters preference group in the side panel. If a higher T_m of probes is required, choose a T_m interval for probes which has higher values than the interval for outer primers.

The output of the design process is a table of solution sets. Each solution set contains the following: a set of primers which are general to all sequences in the alignment, a TaqMan probe which is specific to the set of included sequences (sequences where selection boxes are checked) and a TaqMan probe which is specific to the set of excluded sequences (marked by *). Otherwise, the table is similar to that described above for TaqMan probe prediction on single sequences.

15.10 Analyze primer properties

CLC Gene Workbench 2.0 can calculate and display the properties of predefined primers and probes:

select a primer sequence (primers are represented as DNA sequences in the Navigation Area) | Toolbox in the Menu Bar | Primers and Probes ( ) | Analyze Primer Properties ( )

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the **Project Tree**.

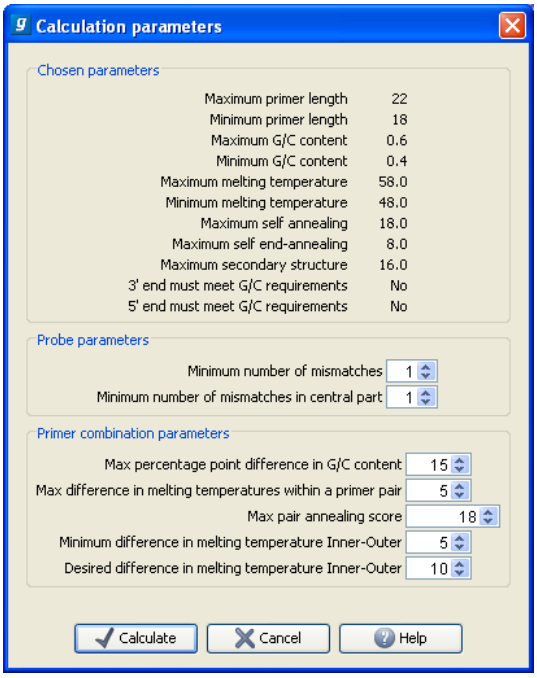

Calculation parameters
Chosen parameters
Maximum primer length 22
Minimum primer length 18
Maximum G/C content 0.6
Minimum G/C content 0.4
Maximum melting temperature 58.0
Minimum melting temperature 48.0
Maximum self annealing 18.0
Maximum self end-annealing 8.0
Maximum secondary structure 16.0
3' end must meet G/C requirements No
5' end must meet G/C requirements No
Probe parameters
Minimum number of mismatches 1
Minimum number of mismatches in central part 1
Primer combination parameters
Max percentage point difference in G/C content 15
Max difference in melting temperatures within a primer pair 5
Max pair annealing score 18
Minimum difference in melting temperature Inner-Outer 5
Desired difference in melting temperature Inner-Outer 10
Calculate
Cancel
Help

Figure 15.14: *Calculation dialog shown when designing alignment based TaqMan probes.*

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. The result is shown in figure 15.15:

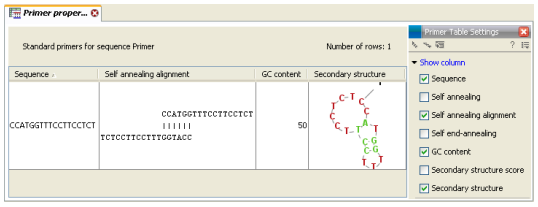


Figure 15.15: *Properties of a primer from the Example Data.*

In the **Side Panel** you can specify the information to display about the primer. The information parameters of the primer properties table are explained in section 15.5.2. The melting temperature of a primer is not available since this requires knowledge of the template to which the primer binds.

15.11 Match primer with sequence

In *CLC Gene Workbench* you have the possibility of matching a known primer against one or more DNA sequences or a list of DNA sequences. This can be applied to test whether a primer used in a previous experiment is applicable to amplify e.g. a homologous region in another species, or to test for potential mispriming.

When applied, the algorithm will search for competing binding sites of the primer within the sequence.

You have the option of choosing the minimum number of matching nucleotides and a minimum number of nucleotides that must bind in the end of the primer. These parameters will be explained in this section.

To search for primer binding sites:

select a nucleotide sequence | **Toolbox in the Menu Bar** | **Primers and Probes (** **)** | **Match Primer with Sequence(** **)**

or **right-click a nucleotide sequence** | **Toolbox** | **Primers and Probes (** **)** | **Match Primer with Sequence (** **)**

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the **Project Tree**.

If you want to adjust the parameters for primer matching, click **Next**.

15.11.1 Search for primer binding sites parameters

This opens the dialog displayed in figure 15.16:

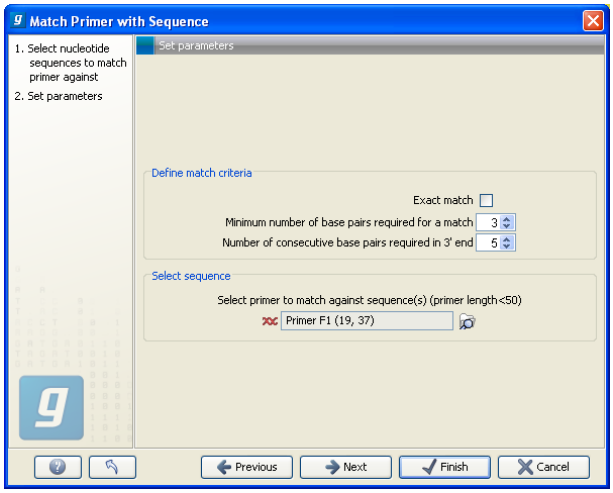


Figure 15.16: *Search parameters for matching primers to the sequence.*

The adjustable parameters for the search are:

- **Exact match**. Choose only to consider exact matches of the primer, i.e. all positions must base pair with the template.
- **Minimum number of base pairs required for a match**. How many nucleotides of the primer that must base pair to the sequence in order to cause priming/mispriming.
- **Number of consecutive base pairs required in 3' end**. How many consecutive 3' end base pairs in the primer that MUST be present to for priming/mispriming to occur. This option is included since 3' terminal base pairs are known to be essential for priming to occur.
- **Select primer to search for**. A primer is a normal DNA sequence, but can only have a maximum length of 50 nucleotides.

After clicking finish the sequences where the primer binds to a subsequence will be annotated with a Primer Binding Site containing information about the primer binding to this subsequence. An example of the result is shown in figure 15.17:

Figure 15.17: *Annotation showing a primer match.*

15.12 Order primers

To facilitate the ordering of primers and probes, *CLC Gene Workbench* offers an easy way of displaying, and saving, a textual representation of one or more primers:

select primers in Navigation Area | Toolbox in the Menu Bar | Primers and Probes ( ) | Order Primers( )

This opens a dialog where you can choose additional primers. Clicking **OK** opens a textual representation of the primers (see 15.18). The first line states the number of primers being ordered and after this follows the names and nucleotide sequences of the primers in 5'-3' orientation. From the editor, the primer information can be copied and pasted to web forms or e-mails. The created object can also be saved to a project and exported as a text file.

See figure 15.18

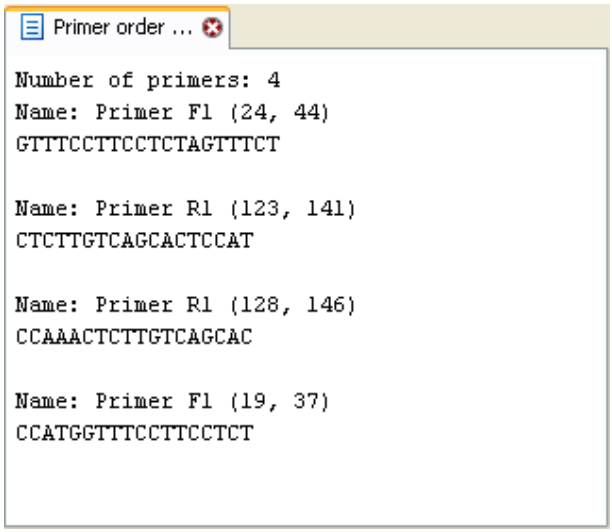


Figure 15.18: *A primer order for 4 primers.*

Chapter 16

Assembly

Contents

16.1 Importing and viewing trace data 208
16.2 Trim sequences 209
16.2.1 Manual trimming 210
16.2.2 Automatic trimming 210
16.3 Assemble sequences 212
16.4 Assemble to reference sequence 213
16.5 Assemble to an existing contig 216
16.6 View and edit contigs 216
16.6.1 Editing and zooming the contig 218
16.6.2 Output from the contig 219
16.6.3 Assembly variance table 219

CLC Gene Workbench 2.0 lets you import, trim and assemble DNA sequence reads from automated sequencing machines. A number of different formats are supported (see section 6.1.1). This chapter first explains how to trim sequence reads. Next follows a description of how to assemble reads into contigs both with and without a reference sequence. In the final section, the options for viewing and editing contigs are explained.

16.1 Importing and viewing trace data

A number of different binary trace data formats can be imported into the program, including *Standard Chromatogram Format (.SCF)*, *ABI sequencer data files (.ABI and .AB1)* and *PHRED output files (.PHD)* (see section 6.1.1).

After import, the sequence reads and their trace data are saved as DNA sequences. This means that all analyzes which apply to DNA sequences can be performed on the sequence reads, including e.g. BLAST and open reading frame prediction.

To view the trace data, open the sequence read in a standard sequence editor. In the Nucleotide info preference group the display of trace data can be selected and unselected. When selected, the trace data information is shown as a plot beneath the sequence. The appearance of the plot can be adjusted using the following options (see figure 16.1):

- **Nucleotide trace.** For each of the four nucleotides the trace data can be selected and unselected.
- **Show confidence.** If confidence information was provided by the base-calling algorithm this can be displayed as a bar plot behind the trace plots. The confidence data is displayed as the log-transformed value of the probability of a given nucleotide position being correctly assigned
- **Show as probabilities.** Displays confidence data as probabilities on a 0-1 scale, i.e. not log-transformed.
- **Scale traces.** A slider which allows the user to scale the trace plots to the desired level of detail.

Figure 16.1: *A sequence with trace data. The preferences for viewing the trace are shown in the Side Panel.*

16.2 Trim sequences

CLC Gene Workbench offers a number of ways to trim your sequence reads prior to assembly. Trimming can be done either as a separate task before assembling, or it can be performed as an integrated part of the assembly process (see section 16.3).

Trimming as a separate task can be done either manually or automatically.

In both instances, trimming of a sequence does not cause data to be deleted, instead both the manual and automatic trimming will put a "Trim" annotation on the trimmed parts as an indication to the assembly algorithm that this part of the data is to be ignored (see figure 16.2). This means that the effect of different trimming schemes can easily be explored without the loss of data. To remove existing trimming from a sequence, simply remove its trim annotation (see section 11.1.4).

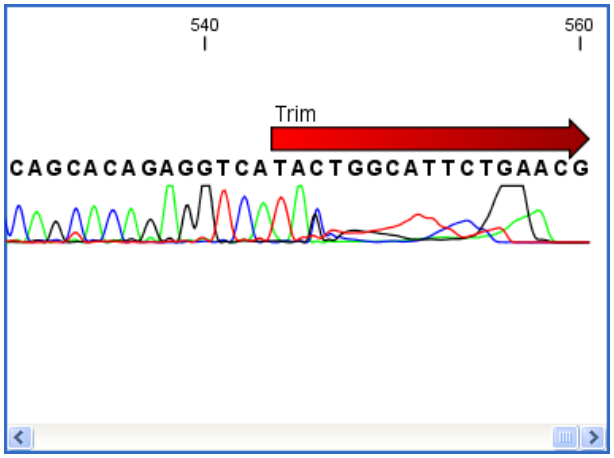


Figure 16.2: *Trimming creates annotations on the regions that will be ignored in the assembly process.*

16.2.1 Manual trimming

Sequence reads can be trimmed manually while inspecting their trace and quality data. Trimming sequences manually corresponds to adding annotation (see also section 11.1.4) but is special in the sense that trimming can only be applied to the ends of a sequence:

> **double-click the sequence to trim in the Navigation Area | enable trace data in the nucleotide info menu group in the preference panel |select the region from where you want trimming to start | right-click the selection | Trim sequence left/right to determine the direction of the trimming**

This will add trimming annotation to the end of the sequence in the selected direction.

16.2.2 Automatic trimming

Sequence reads can be trimmed automatically based on a number of different criteria. Automatic trimming is particularly useful in the following situations:

- If you have many sequence reads to be trimmed.
- If you wish to trim vector contamination from sequence reads.
- If you wish to ensure that the trimming is done according to the same criteria for all the sequence reads.

To trim sequences automatically:

> **select sequence(s) or sequence lists to trim | Toolbox in the Menu Bar | Assembly ( ) | Trim ( )**

This opens a dialog where you can alter your choice of sequences.

When the sequences are selected, click **Next**.

This opens the dialog displayed in figure 16.3.

The following parameters can be adjusted in the dialog:

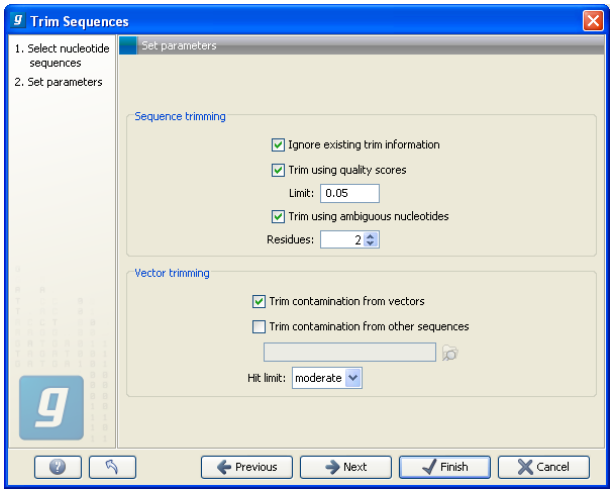


Figure 16.3: *Setting parameters for trimming.*

- **Remove old trimming.** If you have previously trimmed the sequences, you can check this to remove existing trimming annotation prior to analysis.
- **Trim using quality scores.** If the sequence files contain quality scores from a base-caller algorithm this information can be used for trimming sequence ends. The program uses the modified-Mott trimming algorithm for this purpose (Richard Mott personal communication)
- **Trim using ambiguous nucleotides.** This option trims the sequence ends based on the presence of ambiguous nucleotides (typically N). Note that the automated sequencer generating the data must be set to output ambiguous nucleotides in order for this option to apply. The algorithm takes as input the maximal number of ambiguous nucleotides allowed after trimming. If this maximum is set to e.g. 3, the algorithm finds the maximum length region containing 3 or fewer ambiguities and then trims away the ends not included in this region
- **Trim contamination from vectors.** If selected, the program will match the sequence reads against all vectors in the UniVec database and remove sequence ends with significant matches. UniVec is available at `ftp://ftp.ncbi.nih.gov/pub/UniVec/`
- **Trim contamination from other sequences.** This option lets you select a specific vector sequence that you know might be the cause of contamination. Click the **Browse and select object** icon( ) in order to select a sequence.
- **Hit limit.** Specifies how strictly vector contamination is trimmed. Since vector contamination usually occurs at the beginning or end of a sequence, different criteria are applied for terminal and internal matches. A match is considered terminal if it is located within the first 25 bases at either sequence end. Three match categories are defined according to the expected frequency of an alignment with the same score occurring between random sequences, as calculated by NCBI/VecScreen:
  - **Weak.** Expect 1 random match in 40 queries of length 350 kb
    - Terminal match with Score 16 to 18.
    - Internal match with Score 23 to 24.
  - **Moderate.** Expect 1 random match in 1,000 queries of length 350 kb

* Terminal match with Score 19 to 23.
* Internal match with Score 25 to 29.

– **Strong.** Expect 1 random match in 1,000,000 queries of length 350 kb
* Terminal match with Score $\geq$ 24.
* Internal match with Score $\geq$ 30.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. This will start the trimming process. Views of each trimmed sequence will be shown, and you can inspect the result by looking at the "Trim" annotations (they are colored red as default). If there are no trim annotations, the sequence has not been trimmed.

16.3 Assemble sequences

This section describes how to assemble a number of sequence reads into a contig without the use of a reference sequence (a known sequence that can be used for comparison with the other sequences, see section 16.4). To perform the assembly:

select sequences to assemble | Toolbox in the Menu Bar | Assembly ( ) | Assemble Sequences( )

This opens a dialog where you can alter your choice of sequences which you want to assemble. You can also add sequence lists.

When the sequences are selected, click **Next**. This will show the dialog in figure 16.4

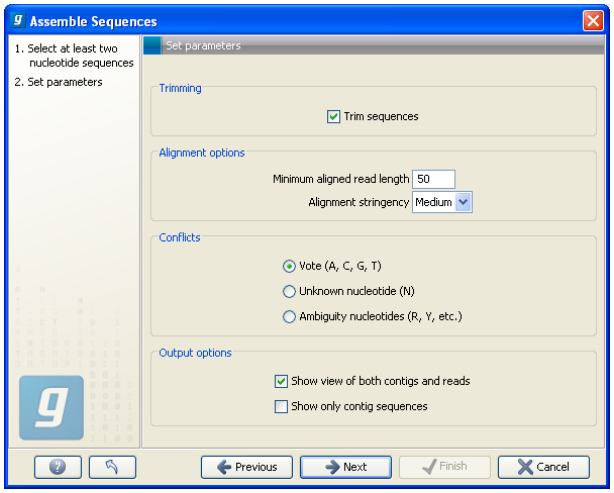


Figure 16.4: *Setting assembly parameters.*

This dialog gives you the following options for assembling:

- **Trim sequences.** If you have not previously trimmed the sequences, this can be done by checking this box. If selected, the next step in the dialog will allow you to specify settings for trimming.
- **Minimum aligned read length.** The minimum number of nucleotides in a read which must be successfully aligned to the contig. If this criteria is not met by a read, this is excluded from the assembly.

- **Alignment stringency.** Specifies the stringency of the scoring function used by the alignment step in the contig assembly algorithm. A higher stringency level will tend to produce contigs with less ambiguities but will also tend to omit more sequencing reads and to generate more and shorter contigs. Three stringency levels can be set:
  - **Low.**
  - **Medium.**
  - **High.**
- **Conflicts.** If there is a conflict between the reads in a given nucleotide position, the program offers three ways to solve this:
  - **Vote (A, C, G, T).** The conflict will be solved by counting instances of each nucleotide and then letting the majority decide the nucleotide in the contig.
  - **Unknown nucleotide (N).** The contig will be assigned an 'N' character in all positions with conflicts.
  - **Ambiguity nucleotides (R, Y, etc.).** The contig will display an ambiguity nucleotide reflecting the different nucleotides found in the reads. For an overview of ambiguity codes, see **Background information** in the **Help** menu.

  Note, that conflicts will always be highlighted no matter which of the options you choose. Furthermore, each conflict will be marked as annotation on the contig sequence and will be present if the contig sequence is extracted for further analysis. As a result, the details of any experimental heterogeneity can be maintained and used when the result of single-sequence analyzes is interpreted.
- **Show view of both contigs and reads.** This will display a proper contig data-object where all the aligned reads are displayed below the contig sequence. (You can always extract the contig sequence without the reads later on.)
- **Show only contig sequences.** This will not display a contig data-object but will only output the assembled contig sequences as single nucleotide sequences. Choosing this option there is thus no opportunity to validate and edit the assembly process.

If you have chosen to "Trim sequences", click **Next** and you will be able to set trim parameters (see section 16.2.2).

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

When the assembly process has ended, a number of views will be shown, each containing a contig of two or more sequences that have been matched. If the number of contigs seem too high or low, try again with another **Alignment stringency** setting. Depending on your choices of output options above, the views will include trace files or only contig sequences. However, the calculation of the contig is carried out the same way, no matter how the contig is displayed.

See section 16.6 on how to use the resulting contigs.

16.4 Assemble to reference sequence

This section describes how to assemble a number of sequence reads into a contig using a reference sequence. A reference sequence can be particularly helpful when the objective is to

characterize SNP variation in the data. Note, that *CLC Gene Workbench* allows you to annotate a reference sequence with known SNP information from the dbSNP database (see section 13.5).

To start the assembly:

select sequences to assemble | Toolbox in the Menu Bar | Assembly ( ) | Assemble Sequences to Reference( )

This opens a dialog where you can alter your choice of sequences which you want to assemble. You can also add sequence lists.

When the sequences are selected, click **Next**, and you will see the dialog shown in figure 16.5

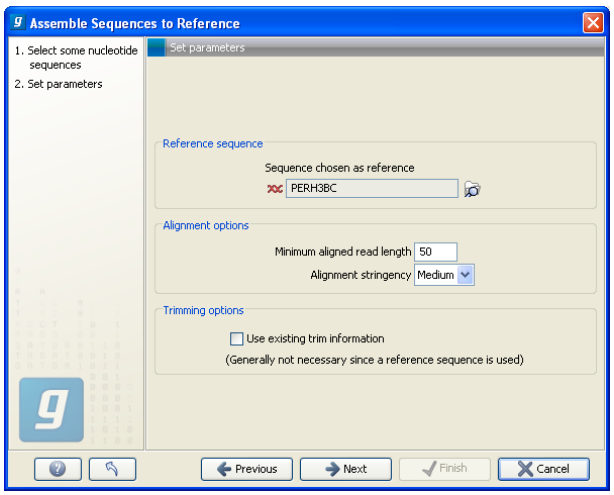


Figure 16.5: *Setting assembly parameters when assembling to a reference sequence.*

This dialog gives you the following options for assembling:

- **Reference sequence.** Click the **Browse and select object** icon( ) in order to select a sequence to use as reference.
- **Minimum aligned read length.** The minimum number of nucleotides in a read which must be successfully aligned to the contig. If this criteria is not met by a read, this is excluded from the assembly.
- **Alignment stringency.** Specifies the stringency of the scoring function used by the alignment step in the contig assembly algorithm. A higher stringency level will tend to produce contigs with less ambiguities but will also tend to omit more sequencing reads and to generate more and shorter contigs. Three stringency levels can be set:
  - **Low.**
  - **Medium.**
  - **High.**
- **Use existing trim information.** When using a reference sequence, trimming is generally not necessary, but if you wish to use trimming you can check this box. It requires that the sequence reads have been trimmed beforehand (see section 16.2 for more information about trimming).

When the parameters have been adjusted, click **Next**, to see the dialog shown in figure 16.6

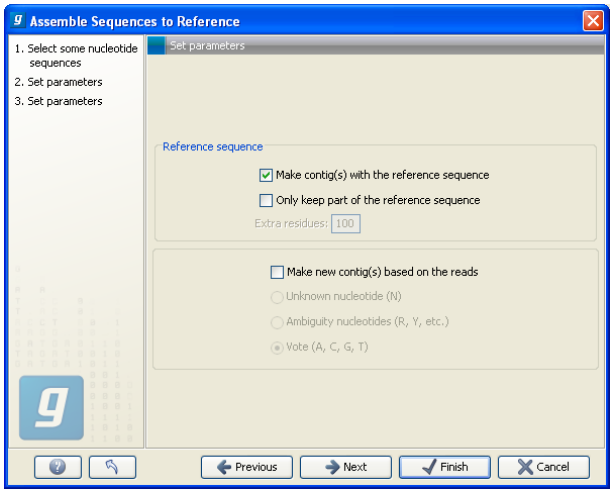


Figure 16.6: *Different options for the output of the assembly.*

In this dialog, you can specify how the result of the assembly - the contig - should be displayed:

- **Make contig(s) with the reference sequence.** This will display a contig data-object with the reference sequence at the top and the reads aligned below. This option is useful when comparing sequence reads to a closely related reference sequence e.g. when sequencing for SNP characterization.
  - **Only keep part of the reference sequence.** If the aligned sequence reads only cover a small part of the reference sequence, it may not be desirable to include the whole reference sequence in the contig data-object. When selected, this option lets you specify how many residues from the reference sequence that should be kept on each side of the region spanned by sequencing reads by entering the number in the **Extra residues** field.
- **Make new contig(s) based on the reads.** This will produce a contig data-object without the reference sequence which resembles that produced when making an ordinary assembly (see section 16.3). In the assembly process the reference sequence is only used as a scaffold for alignment. This option is useful when performing assembly with a reference sequence that is not closely related to the sequencing reads. If there is a conflict between the reads in a given nucleotide position, the program offers three ways to solve this:
  - **Vote (A, C, G, T).** The conflict will be solved by counting instances of each nucleotide and then letting the majority decide the nucleotide in the contig.
  - **Unknown nucleotide (N).** The contig will be assigned an 'N' character in all positions with conflicts.
  - **Ambiguity nucleotides (R, Y, etc.).** The contig will display an ambiguity nucleotide reflecting the different nucleotides found in the reads. For an overview of ambiguity codes, see **Background information** in the **Help** menu.

  Note, that conflicts will always be highlighted no matter which of the options you choose. Furthermore, each conflict will be marked as annotation on the contig sequence and will be present if the contig sequence is extracted for further analysis. As a result, the details of any experimental heterogeneity can be maintained and used when the result of single-sequence analyzes is interpreted.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

16.5 Assemble to an existing contig

This section describes how to assemble sequences to an existing contig. When you assemble to an existing contig, the result of the assembly process is not a new contig where all sequencing reads are re-assembled. Instead, the newly introduced sequencing reads are aligned and added to the existing contig. Sequences that do not align properly to the contig are omitted. This feature can be used for example to provide a steady work-flow when a number of exons from the same gene are sequenced one at a time and assembled to a reference sequence.

To start the assembly:

> **select one contig and a number of sequences | Toolbox in the Menu Bar | Assembly ( ) | Assemble Sequences to Contig( )**

This opens a dialog where you can alter your choice of sequences which you want to assemble. You can also add sequence lists. There has to be one contig among the selected elements.

When the elements are selected, click **Next**, and you will see the dialog shown in figure 16.7

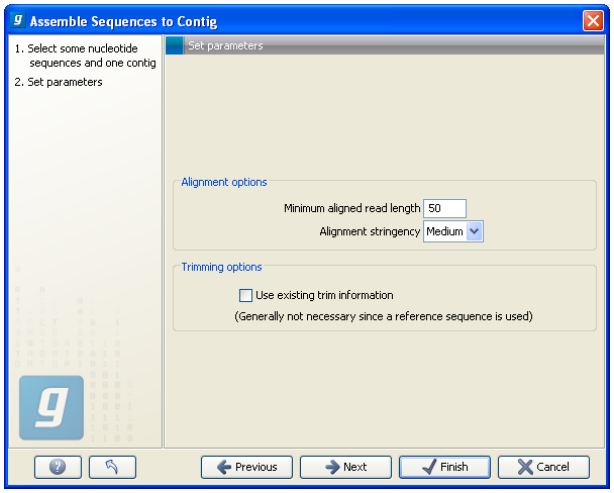


Figure 16.7: *Setting assembly parameters when assembling to an existing contig.*

The options in this dialog are similar to the options that are available when assembling to a reference sequence (see section 16.4).

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. This will start the assembly process.

16.6 View and edit contigs

The result of the assembly process is one or more contigs where the sequence reads have been aligned (see figure 16.8).

You can see that color of the residues and trace at the end of one of the reads has been faded. This indicates, that this region has not contributed to the contig. This may be due to trimming before or during the assembly or due to misalignment to the other reads. Apart from this the view

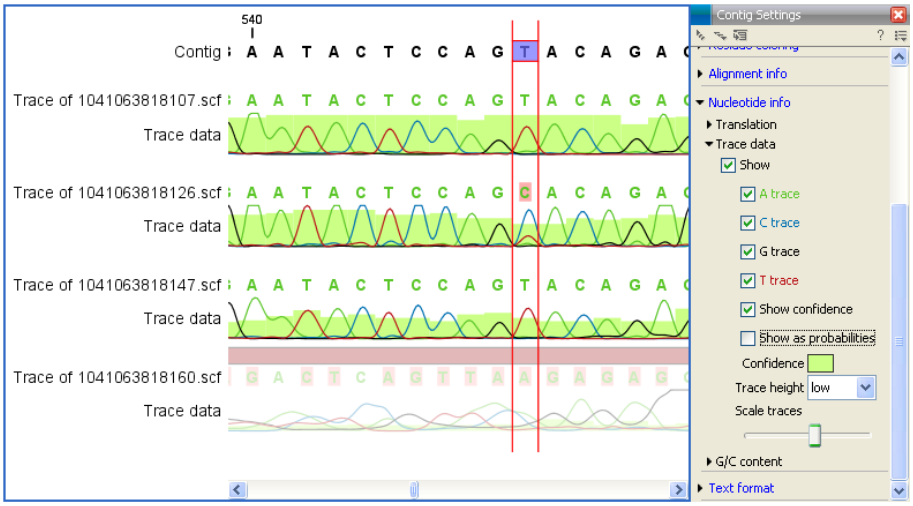


Figure 16.8: *The view of a contig. Notice that you can zoom to a very detailed level in contigs.*

resembles that of alignments (see section 18.2) but has some extra preferences in the **Side Panel**:

- **Assembly Layout.** A new preference group located at the top of the **Side Panel**:
  - **Gather sequences at top**. This option affects the view that is shown when scrolling along a contig. If selected, the sequence reads which did not contribute to the visible part of the contig will be omitted whereas the contributing sequence reads will automatically be placed right below the contig.
  - **Show sequence ends.** Regions that have been trimmed are shown with faded traces and residues. This illustrates that these regions have been ignored during the assembly.
  - **Find Inconsistency**. Clicking this button selects the next position where there is an conflict between the sequence reads. Residues that are different from the contig are colored (as default), providing an overview of the inconsistencies. Since the next inconsistency in the contig is automatically selected it is easy to make changes. You can also use the Space key to find the next inconsistency.
- **Sequence layout.** There is one additional parameter regarding the sequence layout:
  - **Compactness**: In the **Sequence Layout** view preferences, you can control the level of sequence detail to be displayed:
    - **Not compact.** The normal setting with full detail.
    - **Low.** Hides the trace data and puts the reads' annotations on the sequence.
    - **Medium.** The labels of the reads and their annotations are hidden, and the residues of the reads can not be seen.
    - **Compact.** Even less space between the reads.
  - Furthermore, it is not possible to wrap contigs as you can do with alignments.
- **Alignment info.** There is one additional parameter:
  - **Coverage**: Shows how many sequence reads that are contributing information to a given position in the contig. The level of coverage is relative to the overall number of sequence reads that are included in the contig.

        - **Foreground color.** Colors the letters using a gradient, where the left side color is used for low coverage and the right side is used for maximum coverage.
        - **Background color.** Colors the background of the letters using a gradient, where the left side color is used for low coverage and the right side is used for maximum coverage
        - **Graph.** The coverage is displayed as a graph beneath the contig.
            - **Height.** Specifies the height of the graph.
            - **Type.** The graph can be displayed as Line plot, Bar plot or as a Color bar.
            - **Color box.** For Line and Bar plots, the color of the plot can be set by clicking the color box. If a Color bar is chosen, the color box is replaced by a gradient color box as described under Foreground color.

- **Residue coloring.** There is one additional parameter:

    - **Assembly Colors.** This option lets you use different colors for the residues of the contig and the forward and reverse reads. It is particularly useful for getting an overview of forward and reverse reads in the contig:
        - **Contig color.** Colors the residues of the contig sequence with the specified color (can be changed by clicking the colored box).
        - **Forward color.** Colors the residues of forward reads with the specified color (can be changed by clicking the colored box).
        - **Reverse color.** Colors the residues of reverse reads with the specified color (can be changed by clicking the colored box).

Beside from these preferences, all the functionalities of the alignment view are available. This means that you can e.g. add annotations (such as SNP annotation) to regions of interest in the contig.

However, some of the parameters from alignment views are set at a different default value in the view of contigs. Trace data of the sequencing reads are shown if present (can be enabled and disabled under the Nucleotide info preference group), and the **Color different residues** option is also enabled in order to provide a better overview of conflicts (can be changed in the Alignment info preference group).

16.6.1 Editing and zooming the contig

When editing contigs, you are typically interested in confirming or changing single bases, and this can be done simply by selecting the base and typing the right base. Some users prefer to use lower-case letters in order to be able to see which bases were altered when they use the contig later on. In *CLC Gene Workbench* all changes to the contig are recorded in its history log (see section 7) allowing the user to quickly reconstruct the actions performed in the editing session.

There are three shortcut keys for quick editing:

- Space bar: Finds the *next* inconsistency.
- "." (punctuation mark key): Finds the *next* inconsistency.
- "," (comma key): Finds the *previous* inconsistency.

In the contig view, you can use **Zoom in**( ) to zoom to a greater level of detail than in other views (see figure 16.8). This is useful for discerning the trace curves.

If you want to replace a residue with a gap, use the **Delete** or **Backspace** key.

16.6.2 Output from the contig

Due to the integrated nature of *CLC Gene Workbench* it is easy to use the created contig-sequence as input for additional analyzes. If you wish to use the contig-sequence for other analyses:

> **right-click the label of contig (to the left)** | **Open Sequence in New View** | **Save( ) the new sequence**

This will generate a new nucleotide sequence which can be used for e.g. BLAST analysis or cloning construction.

In order to preserve the history of the changes you have made to the contig, the contig itself should be saved from the contig view, using either the save button( ) or by dragging it to the **Navigation Area**.

16.6.3 Assembly variance table

In addition to the standard graphical display of a contig as described above, you can also see a tabular overview of the conflicts in the contig:

> **right-click the tab of the contig** | **Show** | **Table( )**

This will display a new view of the conflicts as shown in figure 16.9.

The table has the following columns:

- **Position.** The position of the conflict measured from the starting point of the contig sequence.
- **Contig Residue.** The contig's residue at this position. The residue can be edited in the graphical view of the contig, as described above.
- **Other Residues.** Lists the residues of the reads. Inside the brackets, you can see the number of reads having this residue at this position. In the example in figure 16.9, you can see that there is a 'C' in the top read in the graphical view. The other two reads have a 'T'. Therefore, the table displays the following text: 'C (1), T (2)'.
- **Note.** Can be used for your own comments on this conflict. Right-click in this cell of the table to add or edit the comments. The comments in the table are associated with the conflict annotation on the contig. Therefore, the comments you enter in the table will also be attached to the annotation on the contig sequence (the comments can be displayed by placing the mouse cursor on the annotation for one second). The comments are saved when you save the contig.

By clicking a row in the table, the corresponding position is highlighted in the graphical view of the contig. Clicking the rows of the table is another way of navigating the contig, apart from using the **Find Inconsistencies** button or using the **Space bar**. You can use the up and down arrow keys to navigate the rows of the table.

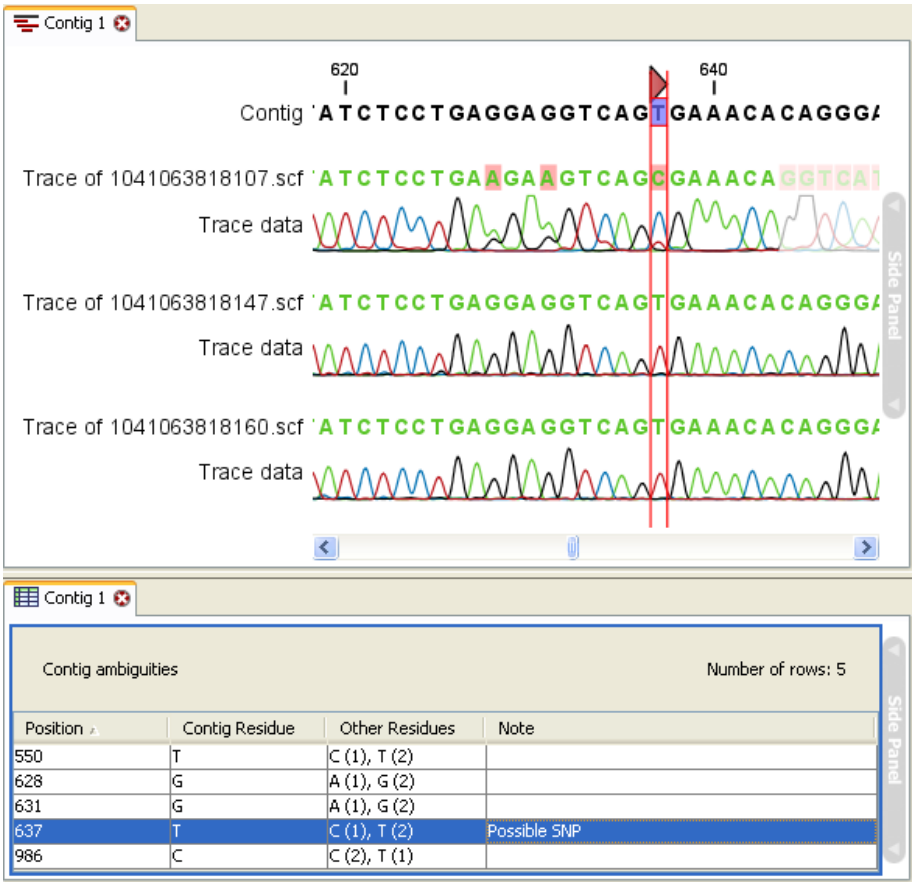


Figure 16.9: *The graphical view of a contig is displayed at the top. At the bottom the conflicts are shown in a table.*

Chapter 17

Cloning and cutting

Contents

17.1 Molecular cloning - an introduction 221
17.2 Graphical display of *in silico* cloning 222
17.2.1 Introduction to the cloning view 222
17.2.2 View preferences for cloning view 223
17.2.3 How to navigate the cloning view 224
17.2.4 Manipulate sequences 225
17.2.5 Insert one sequence into another 228
17.2.6 Show in a circular view 229
17.2.7 Real cloning example 229
17.3 Restriction site analysis 230
17.3.1 Restriction site parameters 230
17.4 Restriction enzyme lists 234
17.4.1 Create enzyme list 234
17.4.2 Modify enzyme list 235
17.5 Gel electrophoresis 235
17.5.1 Separate sequences on gel 236
17.5.2 Separate fragments of sequences using restriction enzymes 237
17.5.3 Gel view 237

CLC Gene Workbench offers graphically advanced *in silico* cloning and design of vectors for various purposes together with restriction enzyme analysis and functionalities for managing lists of restriction enzymes.

First, after a brief introduction, the cloning and vector design is explained. Next, the restriction site analyses are described.

17.1 Molecular cloning - an introduction

Molecular cloning is a very important tool in the quest to understand gene function and regulation. Through molecular cloning it is possible to study individual genes in a controlled environment.

Using molecular cloning is it possible to build complete libraries of fragments of DNA inserted into appropriate cloning vectors.

We offer a significantly different approach for visual cloning than other software tools. In *CLC Gene Workbench* the user is in total control of the cloning process.

17.2 Graphical display of *in silico* cloning

The *in silico* cloning process in *CLC Gene Workbench* begins with the selection of sequences to be used (typically a vector sequence and an insert):

select the sequences in the Navigation Area | Toolbox in the Menu Bar | Cloning and Restriction Sites ( ) | Cloning( )

This will open a view of the selected sequences similar to figure 17.1.

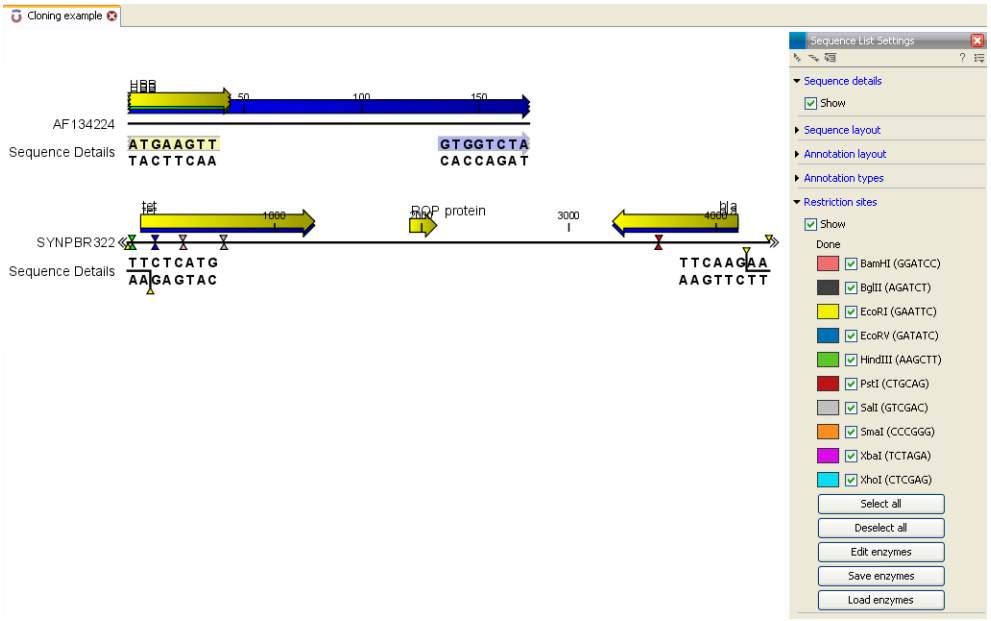


Figure 17.1: *Two sequences in the cloning view.*

If you during the virtual cloning encounter that you need additional sequences, you can easily add more sequences to the view. Just:

right-click anywhere on the empty white area | Add Sequences

17.2.1 Introduction to the cloning view

The cloning view operates with a linear representation of the sequences even though they might be circular. Circular sequence are represented with a small << and >> at the ends of each sequence. When you have finished designing your cloning sequence, you can open it in a circular view (see section 17.2.6).

In the cloning view, most of the basic options for viewing, selecting and zooming the sequences are the same as for the standard sequence viewer. See section 11.1 for an explanation of these options. This means that features such as e.g. known SNP's, exons and other annotations can be displayed on the sequences to guide the choice of regions to clone.

However, the cloning view has many additional interaction possibilities compared to the normal sequence view, and there are several extra visual aids to help you manipulate the sequences. All of this is described in the following.

17.2.2 View preferences for cloning view

Two additional subgroups are shown in the **Side Panel** of the cloning view (see figure 17.2):

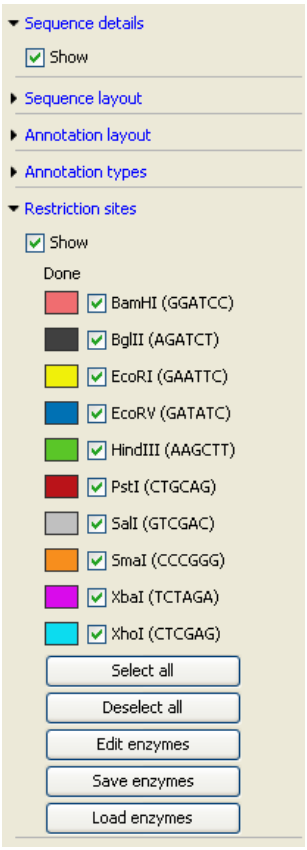


Figure 17.2: *The view preferences for cloning view.*

Sequence details

When you make a selection on the sequence, you will see details of the residues and restriction sites as illustrated in figure 17.3.

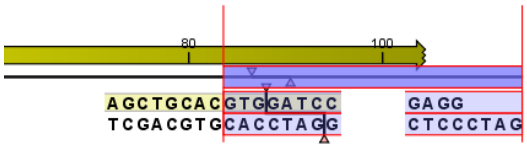


Figure 17.3: *Sequence details of a selection. At the top, the sequence is zoomed out and represented as a black line with annotations, and below the residues are shown (double stranded) with detailed visualization of restriction sites.*

The **Sequence details** are particularly useful when the sequences have overhangs, as shown at the right-side end of the sequence in figure 17.3 which has a "CTAG" overhang.

If you have not made a selection, the details of the ends of the sequences will automatically be shown.

Restriction sites

These preferences allow you to display restriction sites on the sequence. There is a list of enzymes which are represented by different colors. By selecting or deselecting the enzymes in the list, you can specify which enzymes' restriction sites should be displayed (see figure 17.4).

Figure 17.4: *Showing restriction sites of two restriction enzymes.*

The color of the flag of the restriction site can be changed by clicking the colored box next to the enzyme's name.

The list of restriction enzymes contains per default ten of the most popular enzymes, but you can easily modify this list and add more enzymes. You have four ways of modifying the list:

- **Edit enzymes button.** This displays a dialog with the enzymes currently in the list shown at the bottom and a list of available enzymes at the top. To add more enzymes, select them in the upper list and press the **Add enzymes button** ( ). To remove enzymes, select them in the list below and click the **Remove enzymes button** ( ).
- **Load enzymes button.** If you have previously created an enzyme list, you can select this list by clicking the Load enzymes button. You can filter the enzymes in the same way as illustrated in figure 17.13.
- **Add enzymes cutting the selection to panel.** If you make a selection on the sequence, right-click, you find this option for adding enzymes. Based on the entire list of available enzymes, the enzymes cutting in the region you selected will be added to the list in the **Side Panel**.
- **Insert restriction site before/after selection.** If you make a selection on the sequence, right-click, you find this option for inserting a restriction site before or after the region you selected. A dialog is shown where you can select an enzyme whose recognition sequence is inserted. If it was not already present in the list in the **Side Panel**, the enzyme will now be added and selected.

Finally, if you have selected a set of enzymes that you wish to keep for later use, you can click **Save enzymes** and the selected enzymes will be saved to en enzyme list. This list can then be used both when finding restriction sites from the **Toolbox** or when viewing another sequence.

17.2.3 How to navigate the cloning view

The zoom function in the cloning view works on the individual sequence and not the entire view. In that way you can show a long plasmid and short sequence fragments in the same view. However, **Fit Width**( ) and **Zoom to 100%**( ) apply to all the sequences in the view, and can thus be used to reset different zoom levels of the individual sequences.

17.2.4 Manipulate sequences

All manipulations of sequences are done manually, giving you full control over how the sequence is constructed. Manipulations are done through right-click menus which have three different appearances depending on where you click:

- **Right-click the name of the sequence (to the left)**.
- **Right-click a selection**.
- **Right-click a restriction site**.

The three menus are described in the following:

Manipulate the whole sequence

Right-clicking the sequence name at the left side of the view reveals several options on sorting, opening and editing the sequences in the view (see figure 17.5).

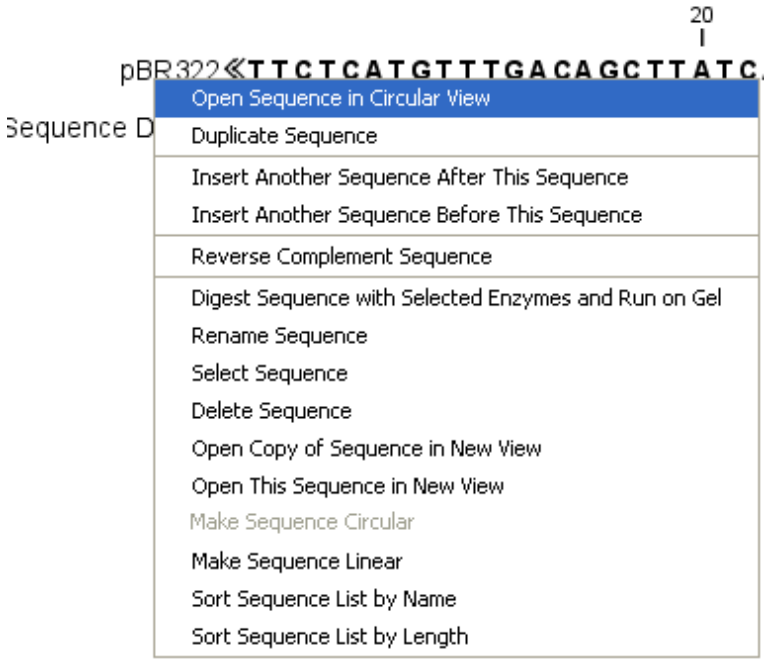


Figure 17.5: *Right click on the sequence in the cloning view.*

- **Insert sequence after this sequence**
  Insert another sequence after this sequence. The sequence to be inserted can be selected from a list which contains the sequences which are present in the cloning editor.
- **Insert sequence before this sequence**
  Insert another sequence before this sequence. The sequence to be inserted can be selected from a list which contains the sequences which are present in the cloning editor.
- **Duplicate sequence**
  Adds a duplicate of the selected sequence. The new sequence will be added to the list of sequences shown on the screen.

- **Reverse complement sequence**
  Creates the reverse complement of a sequence and replaces the original sequence in the list.
- **Make sequence circular**
  This will convert a sequence from a linear to a circular form and has implications for e.g. the action of restriction enzymes. The circular form is represented by << at the beginning of the sequence.
- **Make sequence linear**
  This will convert a sequence from a circular to a linear form.
- **Delete sequence**
  This deletes the given sequence from the cloning editor.
- **Select sequence**
  This will select the entire sequence.
- **Open sequence in new view**
  This will open the selected sequence in a normal sequence view.
- **Sort sequence list by name**
  This will sort all the sequences in the cloning editor alphabetically by name.
- **Sort sequences by length**
  This will sort all the sequences in the cloning editor alphabetically by length.

Manipulate parts of the sequence

Right-clicking a selection reveals several options on manipulating the selection (see figure 17.6).

- **Delete selection**
  This will delete the selected region of the sequence.
- **Duplicate selection**
  If a selection on the sequence is duplicated, the selected region will be added as a new sequence to the cloning editor with a new sequence label representing the length of the fragment.
- **Insert sequence after selection**
  Insert a sequence after the selected region. The sequence to be inserted can be selected from a list containing all sequences in the cloning editor.
- **Insert sequence before selection**
  Insert a sequence before the selected region. The sequence to be inserted can be selected from a list containing all sequences in the cloning editor.
- **Replace selection with sequence**
  This will replace the selected region with a sequence. The sequence to be inserted can be selected from a list containing all sequences in the cloning editor.
- **Cut sequence before selection**
  This will cleave the sequence before the selection and will result in two smaller fragments.

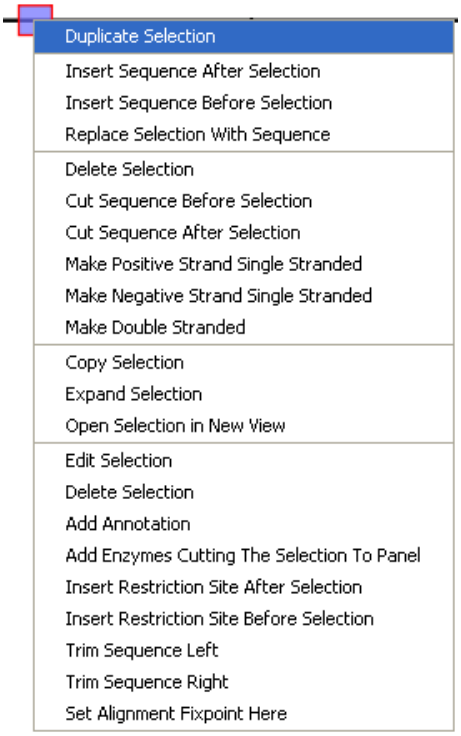


Figure 17.6: *Right click on a sequence selection in the cloning view.*

- **Cut sequence after selection**
  This will cleave the sequence after the selection and will result in two smaller fragments.
- **Copy selection**
  This will copy the selected region to the clipboard, which will enable it for use in other programs.
- **Expand selection**
  This will provide a dialog box in which it is possible to manually expand the selection.
- **Open selection in new view**
  This will open the selected region in the normal sequence view.
- **Edit selection**
  This will open a dialog box, in which is it possible to edit the selected residues.
- **Add annotation**
  This will open the **Add annotation** dialog box.
- **Trim sequences left**
  Adds trim annotation from the beginning of the sequence to the point of selection. Trimmed regions are not included when sequences are assembled into contigs.
- **Trim sequences right**
  Adds trim annotation from the point of selection to the end of the sequence. Trimmed regions are not included when sequences are assembled into contigs.

Manipulate using restriction sites

Right-click on a restriction site (see section 17.2.2) gives you the following options (see figure 17.7):

'Restriction site' in the list below indicates a name on a selection restriction site. This could for example be EcoRV.

- **Insert sequence at this 'Restriction enzyme' site**
  This will insert a sequence from a list into this particular site.
- **Cut this sequence at this 'Restriction enzyme' site**
  This will cut the sequence at this particular site and only this site.
- **Cut this sequence at all 'Restriction enzyme' sites**
  This will cut the sequence at all identical restriction sites, but at no other sites.
- **Cut all sequences at all 'Restriction enzyme' sites**
  This will cut all sequences in the cloning editor view with that particular restriction enzyme. This can potentially generate a lot of sequence fragments.

When a restriction site is double-clicked the recognition site is marked on the sequence and the cut this is marked by arrows.

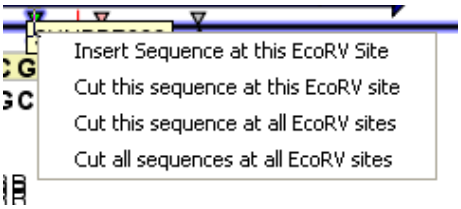

Insert Sequence at this EcoRV Site
Cut this sequence at this EcoRV site
Cut this sequence at all EcoRV sites
Cut all sequences at all EcoRV sites

Figure 17.7: *Right-click on a restriction enzyme annotation in the cloning view.*

When a sequence region between two restriction sites are double-clicked the entire region will automatically be selected.

17.2.5 Insert one sequence into another

Sequences can be inserted into each other in several ways as described above. When you chose to insert one sequence into another you will be presented with a dialog where all sequences in the view are present (see figure 17.8). Select the sequence you want to insert and press ok. If the ends do not fit, a warning will be shown.

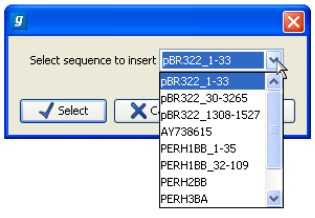


Figure 17.8: *Select a sequence for insertion.*

When the sequence is inserted it will be marked with a selection.

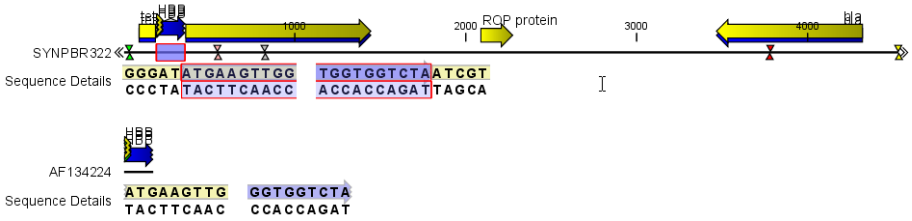


Figure 17.9: *One sequence is now inserted into the cloning vector. The sequence inserted is selected by default.*

17.2.6 Show in a circular view

The sequences stored in the cloning view can be saved to a sequence list and later be opened again for further editing. A sequence list is represented by the following icon in the navigation area ( ).

After finishing the *in silico* cloning in a linear mode the newly formed cloning vector or plasmid can easily be visualized in circular mode. Simply verify that the molecule is circular and right-click the sequence name and press "open sequence in circular view". Then you have a circular view as displayed in figure 17.10.

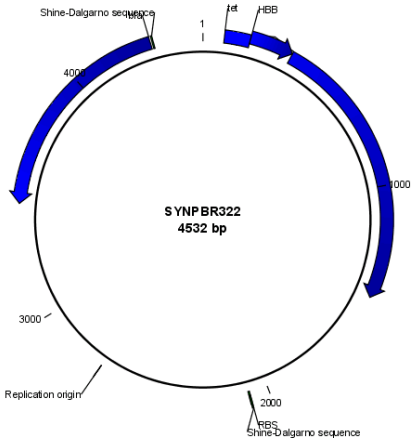


Figure 17.10: *Final circular view of the plasmid. The tetracycline gene is disrupted by an insert.*

17.2.7 Real cloning example

This will show you, very briefly, how to insert a gene into a vector with only a few mouse clicks.

We want to insert a gene Human beta globin 2 (*HBG2* into the commonly known pBR322 plasmid. We choose to insert our gene of interest into the tetracyclin resistance gene of pBR322, which will enable us to select for tetracyclic sensitive clones.

Select the pBR322 and DNA sequence (HUMHBB) holding the gene of interest in the navigator.

Cut out the gene of interest simply by double clicking the gene (*HBG2*), right click with the mouse on the selected region on the sequence and click **duplicate** sequence. Then the selected region is duplicated and shown in the cloning editor as the third sequence. This is seen in figure 17.11.

Now find and select the EcoRV restriction site in position 187 in the donor plasmid, pBR322.

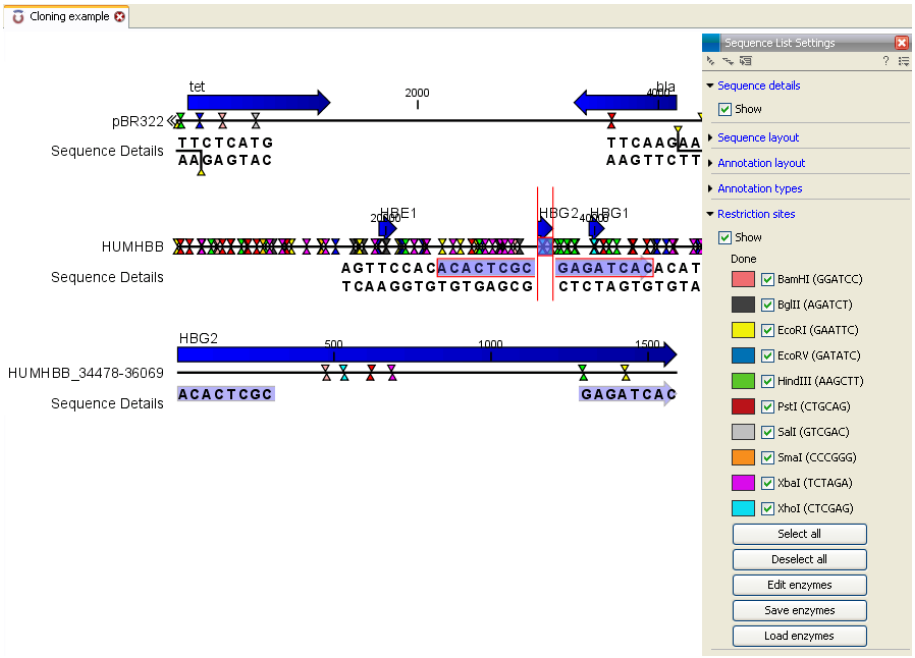


Figure 17.11: *Three sequences are shown in the cloning view. The plasmid, a chromosome holding the gene of interest and the duplicated gene..*

Right click with the mouse on the EcoRV site and click 'insert sequence at this EcoRV site'. Select the third and duplicated sequence, holding the gene of interest, and click **ok**. Now the *HBG2* gene is inserted into the tetracycline gene. Open the sequence in a circular view and see that the tetracycline gene is disrupted by an insert of the *HBG2* gene.

This very short walk-through show some of the powerful cloning capabilities which is included in *CLC Gene Workbench*.

17.3 Restriction site analysis

This section explains how to adjust the detection parameters and offers basic information with respect to restriction site algorithms.

17.3.1 Restriction site parameters

Given a DNA sequence, *CLC Gene Workbench 2.0* detects restriction sites in accordance with detection parameters and shows the detected sites as annotations on the sequence or in textual format in a table.

To detect restriction sites:

select sequence | **Toolbox in the Menu Bar** | **Cloning and Restriction Sites ( )** | **Restriction sites ()**

or **right-click sequence** | **Toolbox** | **Cloning and Restriction Sites ( )** | **Restriction sites ( )**

The result of these steps can be seen in figure 17.12.

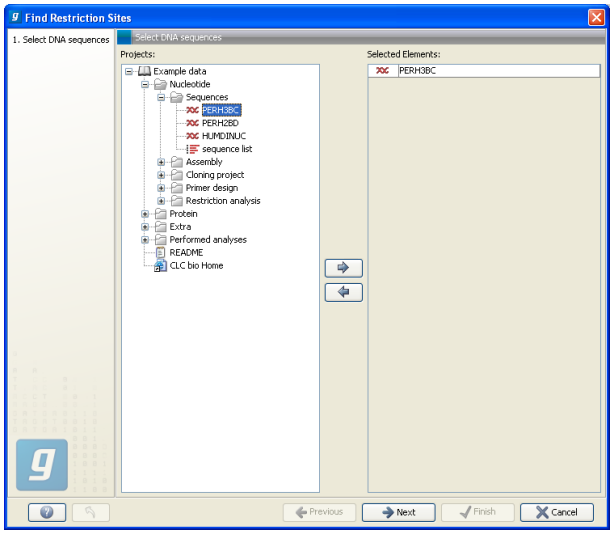


Figure 17.12: *Choosing sequence PERH3BC.*

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the **Project Tree**.

Clicking **Next** generates the dialog shown in figure 17.13.

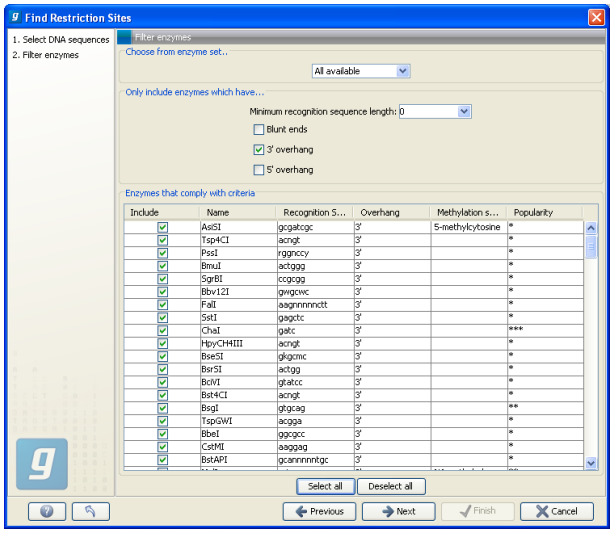


Figure 17.13: *Selecting enzymes.*

In **Step 2** you can adjust which enzymes to use. **Choose from enzyme set...**, allows you to select an enzyme list which is stored in the **Navigation Area**. See section 17.4 for more about creating and modifying enzyme lists.

Only include enzymes which have.... In this part of the dialog, you can limit the number of enzymes included in the list below. You can choose a minimum length of the recognition sequence, and you can choose whether to include enzymes with Blunt ends, 3' overhang, and/or 5' overhang.

Having adjusted the parameters in **Choose from enzyme set...** and **Only include enzymes which have...** the total list of enzymes is shown in the table. The enzymes can be sorted by clicking the column headings, and you can select which enzymes to include in the search be inserting /

removing check marks next to the enzymes.

Clicking **Next** confirms the list of enzymes which will be included in the analysis, and takes you to **Step 3**.

In **Step 3** you can limit which enzymes' cut sites should be included in the output. See figure 17.14.

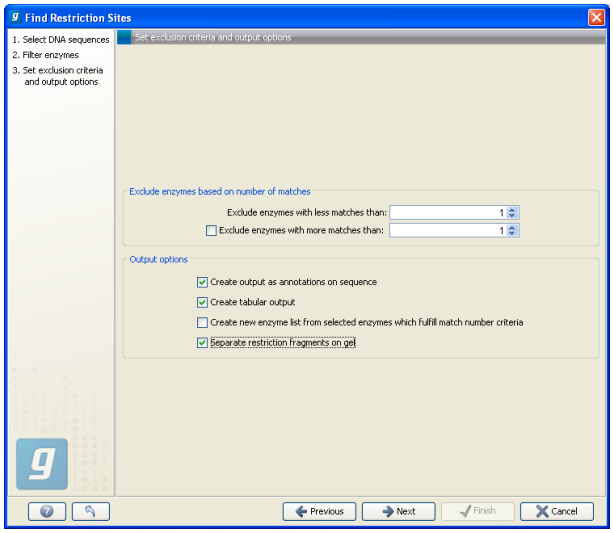

Find Restriction Sites
1. Select DNA sequences
2. Filter enzymes
3. Set exclusion criteria and output options
Set exclusion criteria and output options
Exclude enzymes based on number of matches
Exclude enzymes with less matches than:
Exclude enzymes with more matches than:
Output options
Create output as annotations on sequence
Create tabular output
Create new enzyme list from selected enzymes which fulfill match number criteria
Separate restriction fragments on gel
Previous
Next
Finish
Cancel

Figure 17.14: *Exclusion criteria and output options.*

The default setting **Exclude enzymes with less than 1** (matches), means that enzymes which do not match at all, are not included in the output. If e.g. you only want to see enzymes, which match exactly once, you can check the **Exclude enzymes with more than 1**.

The remaining options relate to the output of the analysis:

- Create output as annotations on sequence
- Create text output
- Create new enzyme list from selected enzymes which fulfill match number criteria
- Separate restriction fragments on gel

If you select the last output option (Separate restriction fragments on gel), there will be one more step. If you have chosen this option, click **Next** to see the dialog shown in figure 17.15.

Here you have four different ways of simulating a gel electrophoresis using the selected restriction enzymes:

- **Cut with selected enzymes and run in one lane.** This will display one lane with a number of bands corresponding to the number of fragments from cutting with the selected enzymes.
- **Cut with selected enzymes and run in one lane per enzyme.** For each of the enzymes selected, there will be a lane displaying the bands of the fragments from cutting just with this enzyme.

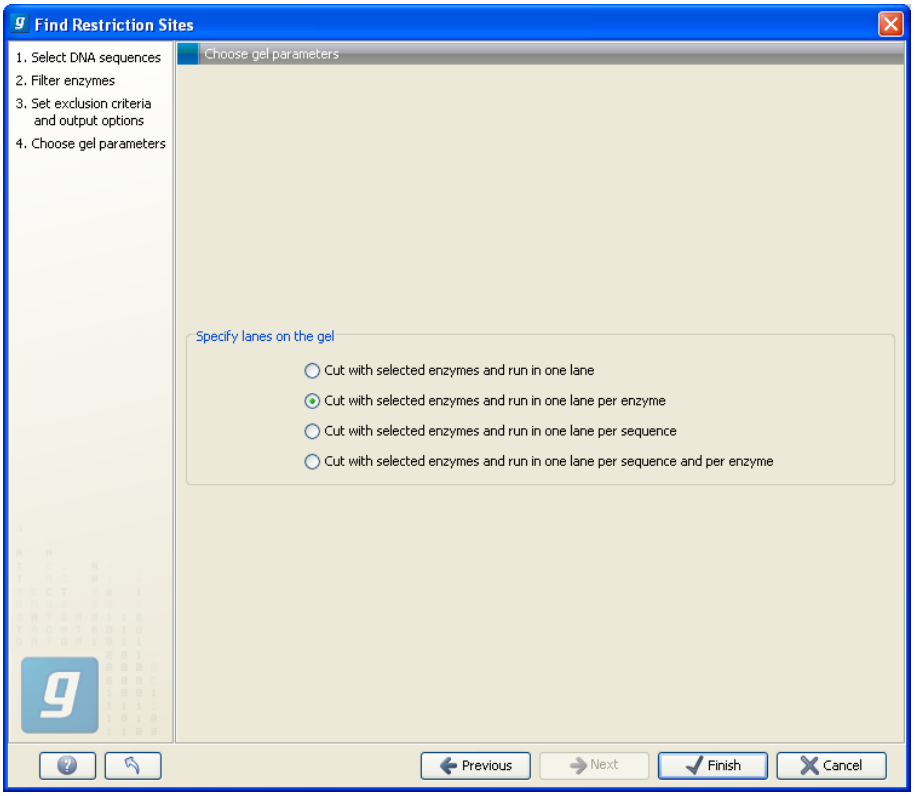

Find Restriction Sites
1. Select DNA sequences
2. Filter enzymes
3. Set exclusion criteria and output options
4. Choose gel parameters
Choose gel parameters
Specify lanes on the gel
Cut with selected enzymes and run in one lane
Cut with selected enzymes and run in one lane per enzyme
Cut with selected enzymes and run in one lane per sequence
Cut with selected enzymes and run in one lane per sequence and per enzyme
Previous
Next
Finish
Cancel

Figure 17.15: *Choosing from four different ways of doing gel electrophoresis.*

- **Cut with selected enzymes and run in one lane per sequence.** If you have selected more than one sequence, this option will display one lane per sequence in the same way as the first option
- **Cut with selected enzymes and run in one lane per sequence and per enzyme.** This will display a number of lanes equalling the number of selected sequences multiplied by the number of selected enzymes, thus combining the functionality of option number two and three.

For more information about gel electrophoresis, see section 17.5.

In order to complete the analysis click **Finish**. The result is shown in figure 17.16.

Choosing the textual output option will open a new view containing a table with an overview of restriction sites. Choosing the graphical output option will add restriction site annotations to the selected sequence.

If too many restriction sites are found, a dialog will ask if you want to proceed or show the restriction sites only in a table format. Showing too many restriction sites as annotations on the sequence will take up a lot of your computer's processing power.

Notice! The text is not automatically saved.

To save the result:

Right-click the tab | **File** | **Save(** **)**

The textual output mentioned above will list all the cut positions where the sequence is restricted.

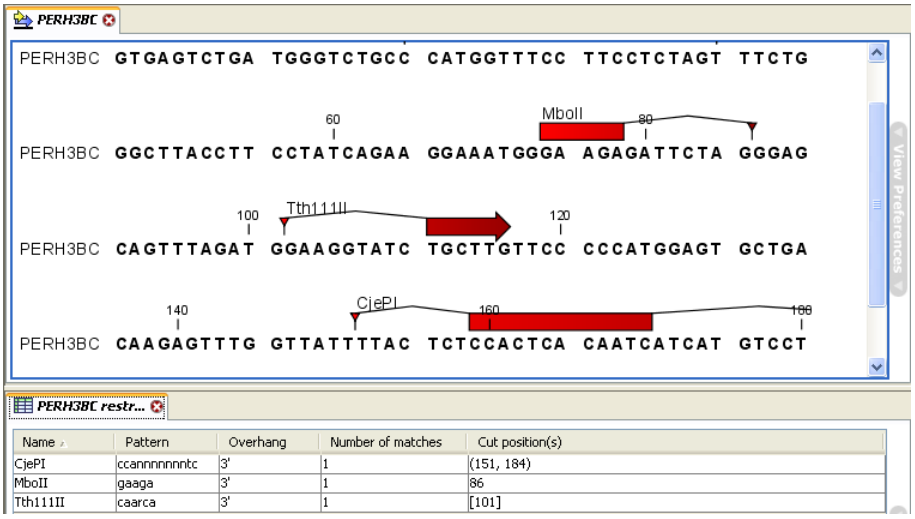


Figure 17.16: *The result of the restriction site detection is displayed as text, and in this example the View Shares the View Area with a View of the PERH3BC sequence displaying the restriction sites (split-screen-view).*

This list may be very long, and hence it might not be possible for *CLC Gene Workbench* to display all cut positions in one cell. If you want to see the entire list of cut positions:

select the table line with the relevant enzyme | **Ctrl + C (⌘ + C on Mac)** | **open a word processing program** | **Ctrl + V (⌘ + V on Mac)**

17.4 Restriction enzyme lists

CLC Gene Workbench includes all the restriction enzymes available in the **REBASE** database. However, when performing restriction site analyses, it is often an advantage to use a customized list of enzymes. In this the user can create special lists containing e.g. all enzymes available in the laboratory freezer, all enzymes used to create a given restriction map or all enzymes that are available form the preferred vendor.
This section describes how you can create an enzyme list, and how you can modify it.

17.4.1 Create enzyme list

CLC Gene Workbench 2.0 uses enzymes from the **REBASE** restriction enzyme database at `http://rebase.neb.com`.
To start creating a sequence list:

right-click in the Navigation Area | **New** | **Enzyme list( )**

This opens the dialog shown in figure 17.17

Step 1 includes two tables. The top table is a list of all the enzymes available in the **REBASE** database. Different information is available for the enzymes, and by clicking the column headings the list can be sorted.
The sequence list is created by adding enzymes to the bottom table. To create sequence list:

Select sequences from top table (hold ctrl (⌘ on Mac)) | **click down-arrow**

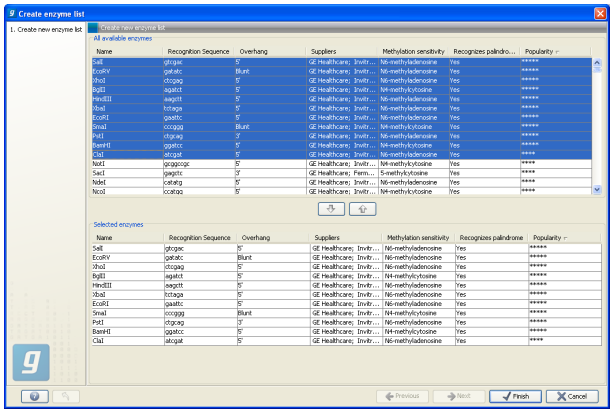


Figure 17.17: *Choosing enzymes for the new enzyme list.*

When the desired enzymes have been chosen, click **Next**.

Choose where to save your enzyme list and name the sequence list. Click **Finish**, to see the enzyme list. In the View preferences it is possible to choose which column to display.

17.4.2 Modify enzyme list

If you want to make changes to an existing enzyme list:

select an enzyme list | Toolbox in the Menu Bar | Cloning and Restriction Sites ( ) | Modify Enzyme List( )

Select the Enzyme list and click **Next**. This opens the dialog shown in figure 17.18.

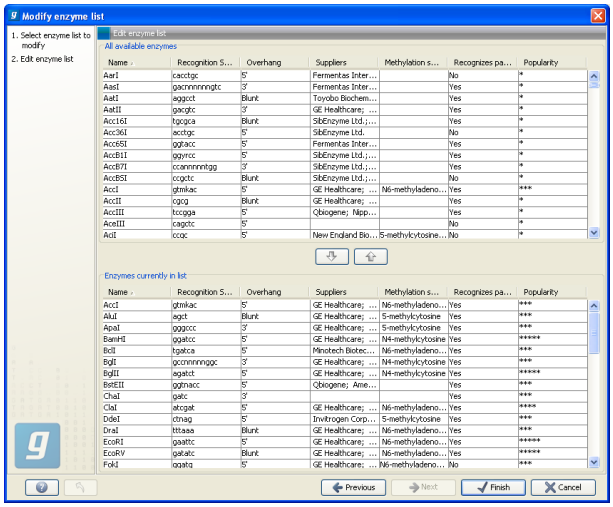


Figure 17.18: *Adding and removing enzymes in the existing enzyme list.*

Select sequences in either top or bottom table (see 17.4.1). Use the arrows to add and remove sequences. Click **Finish** to see the modified list.

17.5 Gel electrophoresis

CLC Gene Workbench enables the user to simulate the separation of nucleotide sequences on a gel. This feature is useful when e.g. designing an experiment which will allow the differentiation

of a successful and an unsuccessful cloning experiment on the basis of a restriction map.

There are two main ways to simulate gel separation of nucleotide sequences:

- A number of existing sequences can be separated on a gel.
- One or more sequences can be digested with restriction enzymes and the resulting fragments can be separated on a gel.

There are several ways to apply these functionalities as described below.

17.5.1 Separate sequences on gel

This section explains how to simulate a gel electrophoresis of one or more existing sequences without restriction enzymes digestion:

select one or more sequences | **Toolbox** | **Cloning and Restriction Sites (** **)** | **Separate Sequences on Gel (** **)**

This opens the dialog shown in figure 17.19.

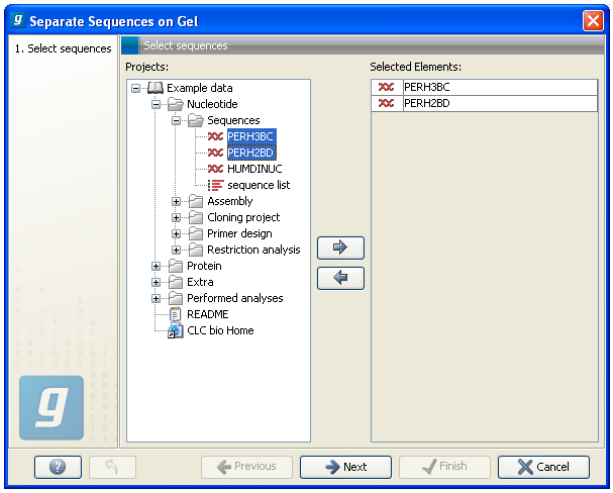


Figure 17.19: *Select one or more sequences to separate on a gel.*

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the **Project Tree**.

Clicking **Next** generates the dialog shown in figure 17.20.

In this dialog, you can choose from two different ways of simulating the gel electrophoresis:

- **Run each sequence in a separate lane.** This will create a new lane for each of the selected sequences. As a result, there will only be one band on each lane.
- **Run all sequences in same lane.** This will create only one lane in which each of the selected sequences will be represented by a band.

The difference between these two options is shown in figure17.21. Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

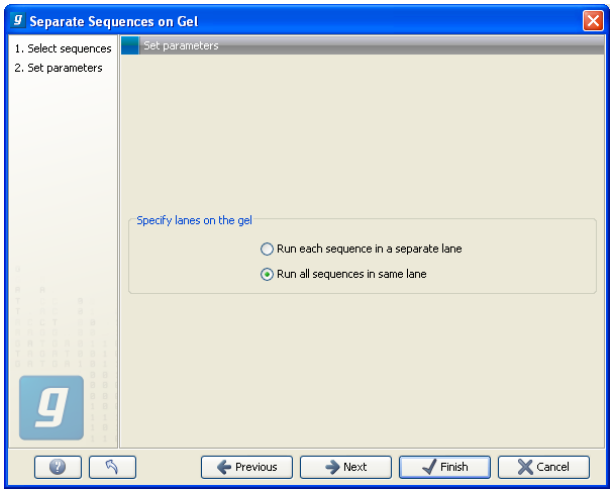


Figure 17.20: *Choosing how to display the lanes.*

For more information about the view of the gel, see section 17.5.3.

17.5.2 Separate fragments of sequences using restriction enzymes

This section explains how to simulate a gel electrophoresis of one or more sequences which are digested with restriction enzymes. There are two ways to do this:

- When performing the **Restriction Sites** analysis from the **Toolbox**, you can choose to separate the resulting fragments on a gel. This is explained in section 17.3.1.
- From all the graphical views of sequences, you can right-click the label of the sequence and choose: **Digest Sequence with Selected Enzymes and Run on Gel**. The views where this option is available are listed below:
  - Circular view (see section 11.6).
  - Ordinary sequence view (see section 11.1).
  - Graphical view of sequence lists (see section 11.5).
  - Cloning editor (see section 17.2).
  - Primer designer (see section 15.3).

  Furthermore, you can also right-click an empty part of the view of the graphical view of sequence lists and the cloning editor and choose **Digest All Sequences with Selected Enzymes and Run on Gel**. This opens a dialog with functionalities similar to the one in figure 17.15.

  Notice! When using the right-click options, the sequence will be digested with the enzymes that are selected in the **Side Panel**. This is explained in section 11.1.1.

17.5.3 Gel view

In figure 17.22 you can see a simulation of a gel with its **Side Panel** to the right. This view will be explained in this section.

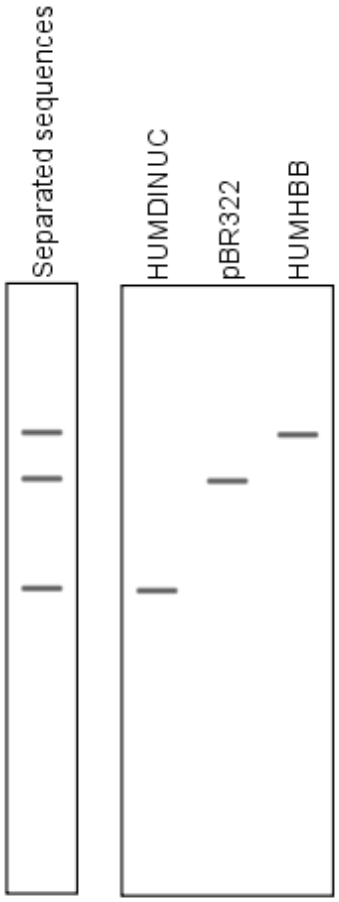


Figure 17.21: *Gel electrophoresis of three sequences. The left side shows the sequences together in one lane, each represented by a band. The right side shows a lane for each sequence.*

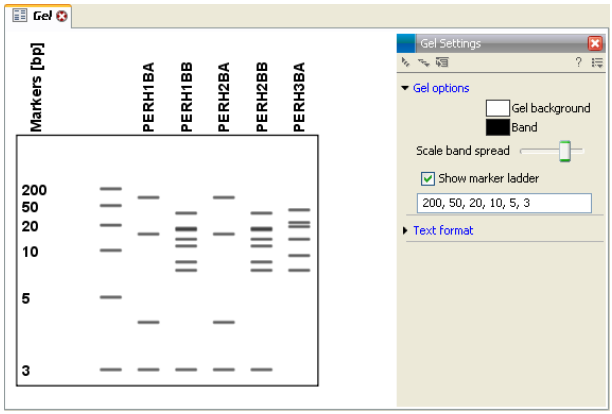


Figure 17.22: *Five lanes showing fragments of five sequences cut with restriction enzymes.*

Information on bands and fragment size

You can get information about the individual bands by hovering the mouse cursor on the band of interest. This will display a tool tip with information about the fragment size, and for lanes comparing whole sequences, you will also see the sequence name.

Notice! You have to be in **Selection** ( ) or **Pan** ( ) mode in order to get this information.

It can be useful to add markers to the gel which enables you to compare the sizes of the bands. This is done by clicking **Show marker ladder** in the **Side Panel**.

You enter the markers by writing them in the text field, separated by commas.

Modifying the layout

The background of the lane and the colors of the bands can be changed in the **Side Panel**. Click the colored box to display a dialog for picking a color. The slider **Scale band spread** can be used to adjust the effective time of separation on the gel, i.e. how much the bands will be spread over the lane. In a real electrophoresis experiment this property will be determined by several factors including time of separation, voltage and gel density.

You can also modify the layout of the view by zooming in or out. Click **Zoom in** ( ) or **Zoom out** ( ) in the Toolbar and click the view.

Finally, you can modify the format of the text heading each lane in the **Text format** preferences in the **Side Panel**.

Chapter 18

Sequence alignment

Contents

18.1 Create an alignment 241
18.1.1 Gap costs 242
18.1.2 Fast or accurate alignment algorithm 242
18.1.3 Aligning alignments 243
18.1.4 Fixpoints 244
18.2 View alignments 246
18.2.1 Sequence logo 247
18.2.2 Conservation 248
18.2.3 Gap fraction 249
18.3 Edit alignments 249
18.3.1 Move residues and gaps 249
18.3.2 Insert gap columns 249
18.3.3 Delete residues and gaps 250
18.3.4 Copy annotations to other sequences 250
18.3.5 Move sequences up and down 250
18.3.6 Delete and add sequences 250
18.3.7 Realign selection 251
18.4 Join alignments 251
18.4.1 How alignments are joined 252
18.5 Bioinformatics explained: Multiple alignments 253
18.5.1 Use of multiple alignments 253
18.5.2 Constructing multiple alignments 254

CLC Gene Workbench 2.0 can align nucleotides and proteins using a *progressive alignment* algorithm (see section 18.5 or read the White paper on alignments in the **Science** section of `http://www.clcbio.com`).

This chapter describes how to use the program to align sequences. The chapter also describes alignment algorithms in more general terms.

18.1 Create an alignment

Alignments can be created from sequences, sequence lists (see section 11.5), existing alignments and from any combination of the three.

To create an alignment in *CLC Gene Workbench 2.0*:

select elements to align | **Toolbox in the Menu Bar** | **Alignments and Trees( )** | **Create Alignment ( )**

or **select elements to align** | **right-click either selected sequence** | **Toolbox** | **Alignments and Trees( )** | **Create Alignment ( )**

This opens the dialog shown in figure 18.1.

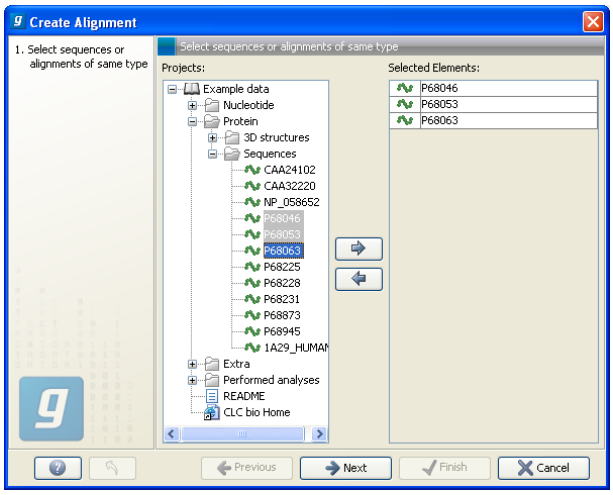


Figure 18.1: *Creating an alignment.*

If you have selected some elements before choosing the Toolbox action, they are now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences, sequence lists or alignments from the **Project Tree**. Click **Next** to adjust alignment algorithm parameters. Clicking **Next** opens the dialog shown in figure 18.2.

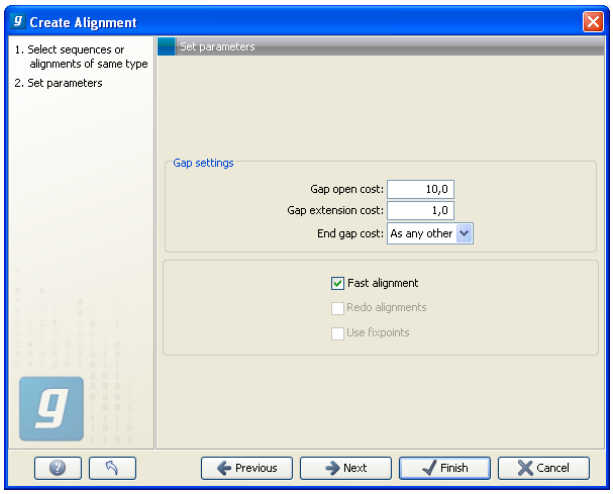


Figure 18.2: *Adjusting alignment algorithm parameters.*

18.1.1 Gap costs

The alignment algorithm has three parameters concerning gap costs: Gap open cost, Gap extension cost and End gap cost. The precision of these parameters is to one place of decimal.

- **Gap open cost**. The price for introducing gaps in an alignment.
- **Gap extension cost**. The price for every extension past the initial gap.

If you expect a lot of small gaps in your alignment, the Gap open cost should equal the Gap extension cost. On the other hand, if you expect few but large gaps, the Gap open cost should be set significantly higher than the Gap extension cost.

However, for most alignments it is a good idea to make the Gap open cost quite a bit higher than the Gap extension cost. The default values are 10.0 and 1.0 for the two parameters, respectively.

- **End gap cost**. The price of gaps at the beginning or the end of the alignment. One of the advantages of the *CLC Gene Workbench 2.0* alignment method is that it provides flexibility in the treatment of gaps at the ends of the sequences. There are three possibilities:
  - **Free end gaps**. Any number of gaps can be inserted in the ends of the sequences without any cost.
  - **Cheap end gaps**. All end gaps are treated as gap extensions and any gaps past 10 are free.
  - **End gaps as any other**. Gaps at the ends of sequences are treated like gaps in any other place in the sequences.

When aligning a long sequence with a short partial sequence, it is ideal to use free end gaps, since this will be the best approximation to the situation. The many gaps inserted at the ends are not due to evolutionary events, but rather to partial data.

Many homologous proteins have quite different ends, often with large insertions or deletions. This confuses alignment algorithms, but using the ''cheap end gaps'' option, large gaps will generally be tolerated at the sequence ends, improving the overall alignment. This is the default setting of the algorithm.

Finally, treating end gaps like any other gaps is the best option when you know that there are no biologically distinct effects at the ends of the sequences.

Figures 18.3 and 18.4 illustrate the differences between the different gap scores at the sequence ends.

18.1.2 Fast or accurate alignment algorithm

CLC Gene Workbench has two algorithms for calculating alignments:

- **Accurate alignment.** This is the recommended choice unless you find the processing time too long.
- **Fast alignment.** This allows for use of an optimized alignment algorithm which is very fast. The fast option is particularly useful for datasets with very long sequences.

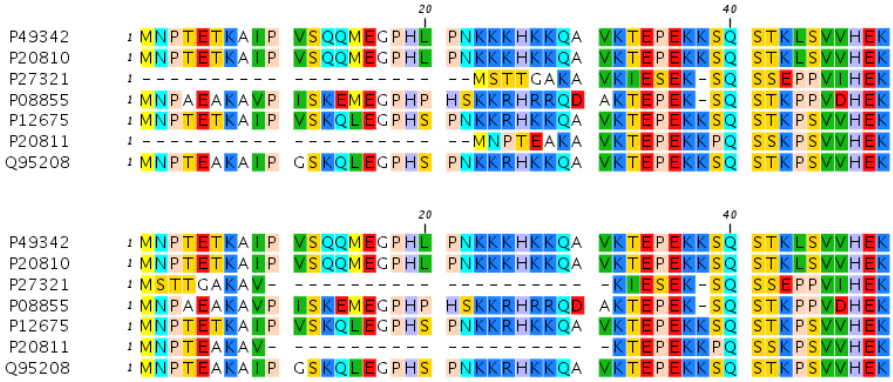

20
40
P49342 1 MNPTETKAIP VSQQMEGPHL PNKKKHKKQA VKTEPEKKSQ STKLSVVHEK
P20810 1 MNPTETKAIP VSQQMEGPHL PNKKKHKKQA VKTEPEKKSQ STKLSVVHEK
P27321 1 ---------- ---------- --MSTTGAKA VKIESEK-SQ SSEPPVIHEK
P08855 1 MNPAEAKAVP ISKEMEGPHP HSKKRHRRQD AKTEPEK-SQ STKPPVDHEK
P12675 1 MNPTETKAIP VSKQLEGPHS PNKKRHKKQA VKTEPEKKSQ STKPSVVHEK
P20811 1 ---------- ---------- --MNPTEAKA VKTEPEKKPQ SSKPSVVHEK
Q95208 1 MNPTEAKAIP GSKQLEGPHS PNKKRHKKQA VKTEPEKKSQ STKPSVVHEK
20
40
P49342 1 MNPTETKAIP VSQQMEGPHL PNKKKHKKQA VKTEPEKKSQ STKLSVVHEK
P20810 1 MNPTETKAIP VSQQMEGPHL PNKKKHKKQA VKTEPEKKSQ STKLSVVHEK
P27321 1 MSTTGAKAV- ---------- ---------- -KIESEK-SQ SSEPPVIHEK
P08855 1 MNPAEAKAVP ISKEMEGPHP HSKKRHRRQD AKTEPEK-SQ STKPPVDHEK
P12675 1 MNPTETKAIP VSKQLEGPHS PNKKRHKKQA VKTEPEKKSQ STKPSVVHEK
P20811 1 MNPTEAKAV- ---------- ---------- -KTEPEKKPQ SSKPSVVHEK
Q95208 1 MNPTEAKAIP GSKQLEGPHS PNKKRHKKQA VKTEPEKKSQ STKPSVVHEK

Figure 18.3: *The first 50 positions of two different alignments of seven calpastatin sequences. The top alignment is made with cheap end gaps, while the bottom alignment is made with end gaps having the same price as any other gaps. In this case it seems that the latter scoring scheme gives the best result.*

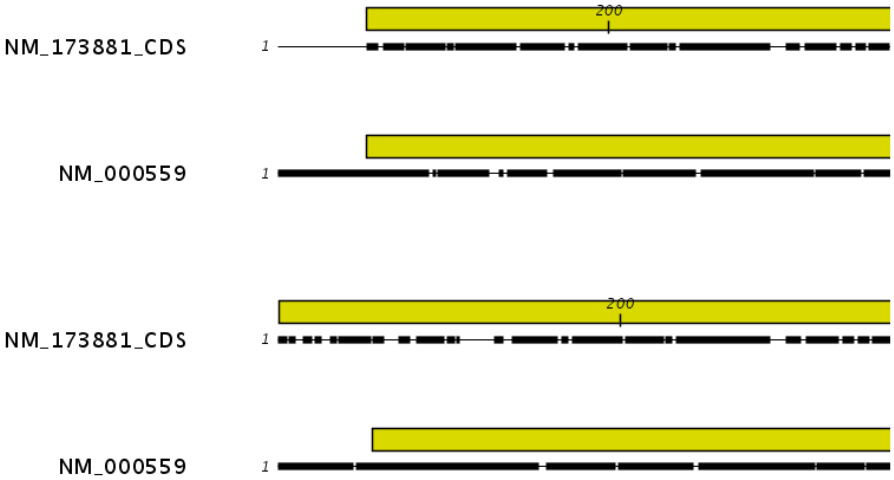


Figure 18.4: *The alignment of the coding sequence of bovine myoglobin with the full mRNA of human gamma globin. The top alignment is made with free end gaps, while the bottom alignment is made with end gaps treated as any other. The yellow annotation is the coding sequence in both sequences. It is evident that free end gaps are ideal in this situation as the start codons are aligned correctly in the top alignment. Treating end gaps as any other gaps in the case of aligning distant homologs where one sequence is partial leads to a spreading out of the short sequence as in the bottom alignment.*

For a comprehensive explanation of the alignment algorithms see section 18.5.

18.1.3 Aligning alignments

If you have selected an existing alignment in the first step (18.1), you have to decide how this alignment should be treated.

- **Redo alignment.** The original alignment will be realigned if this checkbox is checked. Otherwise, the original alignment is kept in its original form except for possible extra equally sized gaps in all sequences of the original alignment. This is visualized in figure 18.5.

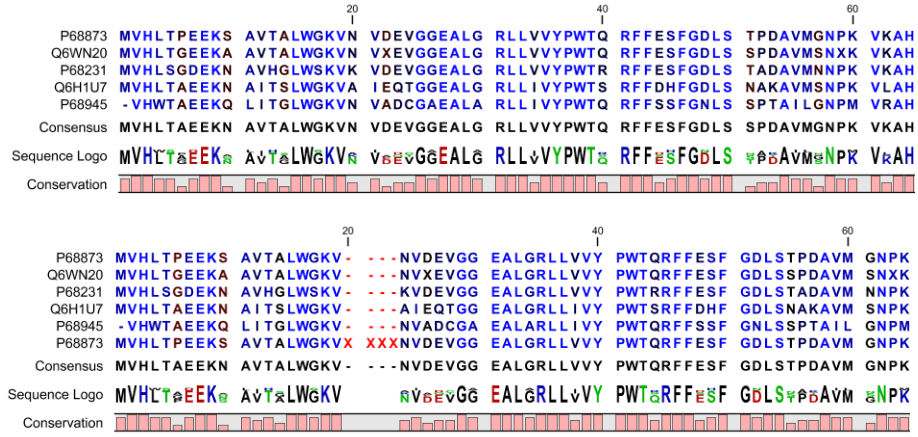

P68873 MVHLTPEEKS AVTALWGKVN VDEVGGEALG RLLVVYPWTQ RFFESFGDLS TPDAVMGNPK VKAH
Q6WN20 MVHLTGEEKA AVTALWGKVN VXEVGGEALG RLLVVYPWTQ RFFESFGDLS SPDAVMSNXK VKAH
P68231 MVHLSGDEKN AVHGLWSKVK VDEVGGEALG RLLVVYPWTR RFFESFGDLS TADAVMNNPK VKAH
Q6H1U7 MVHLTAEEKN AITSLWGKVA IEQTGGEALG RLLIVYPWTS RFFDHFGDLS NAKAVMSNPK VLAH
P68945 -VHWTAEEKQ LITGLWGKVN VADCGAEALA RLLIVYPWTQ RFFSSFGNLS SPTAILGNPM VRAH
Consensus MVHLTAEEKN AVTALWGKVN VDEVGGEALG RLLVVYPWTQ RFFESFGDLS SPDAVMGNPK VKAH
Sequence Logo
Conservation
P68873 MVHLTPEEKS AVTALWGKV- ---NVDEVGG EALGRLLVVY PWTQRFFESF GDLSTPDAVM GNPK
Q6WN20 MVHLTGEEKA AVTALWGKV- ---NVXEVGG EALGRLLVVY PWTQRFFESF GDLSSPDAVM SNXK
P68231 MVHLSGDEKN AVHGLWSKV- ---KVDEVGG EALGRLLVVY PWTRRFFESF GDLSTADAVM NNPK
Q6H1U7 MVHLTAEEKN AITSLWGKV- ---AIEQTGG EALGRLLIVY PWTSRFFDHF GDLSNAKAVM SNPK
P68945 -VHWTAEEKQ LITGLWGKV- ---NVADCGA EALARLLIVY PWTQRFFSSF GNLSSPTAIL GNPM
P68873 MVHLTPEEKS AVTALWGKVX XXXNVDEVGG EALGRLLVVY PWTQRFFESF GDLSTPDAVM GNPK
Consensus MVHLTAEEKN AVTALWGKV- ---NVDEVGG EALGRLLVVY PWTQRFFESF GDLSTPDAVM GNPK
Sequence Logo
Conservation

Figure 18.5: *The top figures shows the original alignment. In the bottom panel a single sequence with four inserted X's are aligned to the original alignment. This introduces gaps in all sequences of the original alignment. All other positions in the original alignment are fixed.*

This feature is useful if you wish to add extra sequences to an existing alignment, in which case you just select the alignment and the extra sequences and choose not to redo the alignment.

It is also useful if you have created an alignment where the gaps are not placed correctly. In this case, you can realign the alignment with different gap cost parameters.

18.1.4 Fixpoints

With fixpoints, you can get full control over the alignment algorithm. The fixpoints are points on the sequences that are forced to align to each other.

Fixpoints are added to sequences or alignments before clicking "Create alignment". To add a fixpoint, open the sequence or alignment and:

> **Select the region you want to use as a fixpoint | right-click the selection | Set alignment fixpoint here**

This will add an annotation labeled "Fixpoint" to the sequence (see figure 18.6). Use this procedure to add fixpoints to the other sequence(s) that should be forced to align to each other.

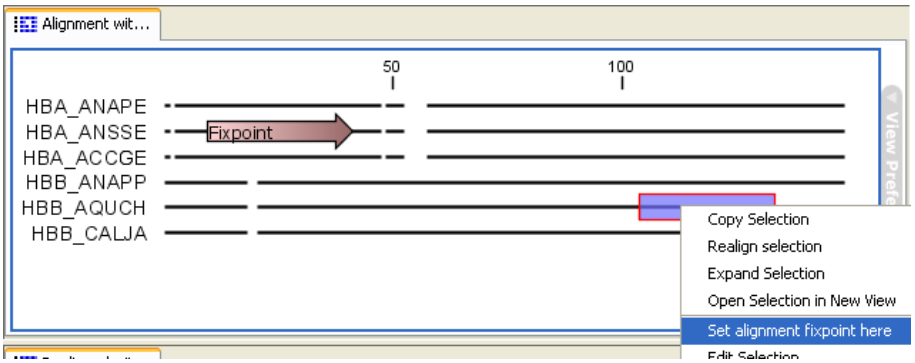


Figure 18.6: *Adding a fixpoint to a sequence in an existing alignment. At the top you can see a fixpoint that has already been added.*

When you click "Create alignment" and go to **Step 2**, check **Use fixpoints** in order to force the

alignment algorithm to align the fixpoints in the selected sequences to each other.

In figure 18.7 the result of an alignment using fixpoints is illustrated.

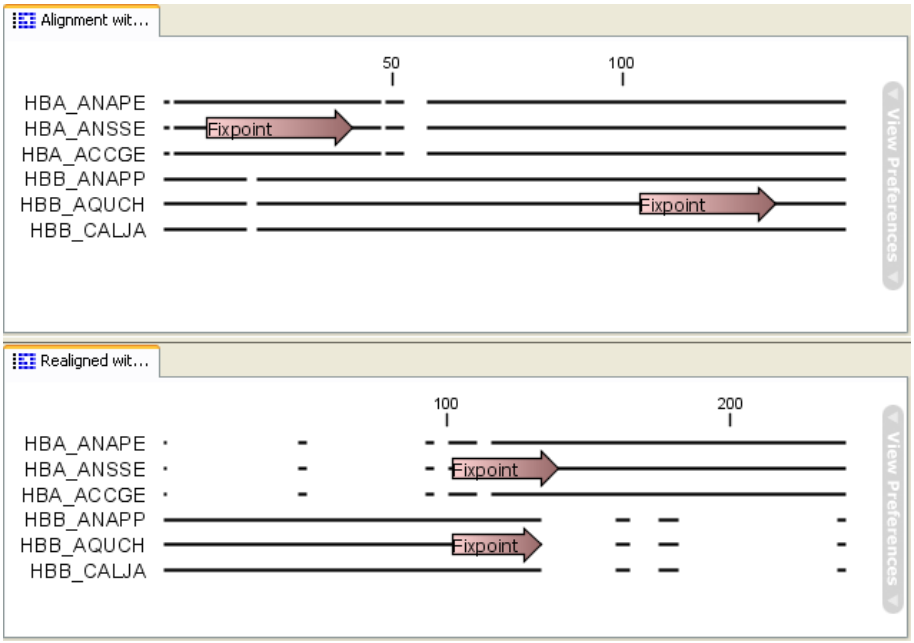


Figure 18.7: *Realigning using fixpoints. In the top view, fixpoints have been added to two of the sequences. In the view below, the alignment has been realigned using the fixpoints. The three top sequences are very similar, and therefore they follow the one sequence (number two from the top) that has a fixpoint.*

You can add multiple fixpoints, e.g. adding two fixpoints to the sequences that are aligned will force their first fixpoints to be aligned to each other, and their second fixpoints will also be aligned to each other.

Advanced use of fixpoints

Fixpoints with the same names will be aligned to each other, which gives the opportunity for great control over the alignment process. It is only necessary to change any fixpoint names in very special cases.

One example would be three sequences A, B and C where sequences A and B has one copy of a domain while sequence C has two copies of the domain. You can now force sequence A to align to the first copy and sequence B to align to the second copy of the domains in sequence C. This is done by inserting fixpoints in sequence C for each domain, and naming them 'fp1' and 'fp2' (for example). Now, you can insert a fixpoint in each of sequences A and B, naming them 'fp1' and 'fp2', respectively. Now, when aligning the three sequences using fixpoints, sequence A will align to the first copy of the domain in sequence C, while sequence B would align to the second copy of the domain in sequence C.

You can name fixpoints by:

> **right-click the Fixpoint annotation | Edit Annotation | type the name in the 'Name' field**

18.2 View alignments

Since an alignment is a display of several sequences arranged in rows, the basic options for viewing alignments are the same as for viewing sequences. Therefore we refer to section 11.1 for an explanation of these basic options.

However, there are a number of alignment-specific view options in the **Alignment info** preference group in the **Side Panel** to the right of the view. These preferences relate to each column in the alignment. Below is more information on these view options.

- **Consensus.** Shows a consensus sequence at the bottom of the alignment. The consensus sequence is based on every single position in the alignment and reflects an artificial sequence which resembles the sequence information of the alignment, but only as one single sequence. If all sequences of the alignment is 100% identical the consensus sequence will be identical to all sequences found in the alignment. If the sequences of the alignment differ the consensus sequence will reflect the most common sequences in the alignment. Parameters for adjusting the consensus sequences are described above.

  The **Consensus Sequence** can be opened in a new view, simply by right-clicking the **Consensus Sequence** and click **Open Consensus in New View**.

  - **Limit.** This option determines how conserved the sequences must be in order to agree on a consensus.
  - **No gaps.** Checking this option will not show gaps in the consensus.
  - **Ambiguous symbol.** Select how ambiguities should be displayed in the consensus line.

- **Sequence logo.**See section 18.2.1 for more details.

  - **Foreground color.** Colors the letters using a gradient according to the information content of the alignment column.
  - **Background color.** Sets a background color of the residues using a gradient in the same way as described above.
  - **Graph.** Displays sequence logo at the bottom of the alignment.
    - **Height.** Specifies the height of the sequence logo graph.
    - **Color.** The sequence logo can be displayed in black or Rasmol colors. For protein alignments, a polarity color scheme is also available, where hydrophobic residues are shown in black color, hydrophilic residues as green, acidic residues as red and basic residues as blue.

- **Conservation.** Displays the level of conservation at each position in the alignment.

  - **Foreground color.** Colors the letters using a gradient, where the right side color is used for highly conserved positions and the left side color is used for positions that are less conserved.
  - **Background color.** Sets a background color of the residues using a gradient in the same way as described above.
  - **Graph.** Displays the conservation level as a graph at the bottom of the alignment. The bar (default view) show the conservation of all sequence positions. The height of the graph reflects how conserved that particular position is in the alignment. If one position is 100% conserved the graph will be shown in full height.

        - **Height.** Specifies the height of the graph.
        - **Type.** The type of the graph.
            - **Line plot.** Displays the graph as a line plot.
            - **Bar plot.** Displays the graph as a bar plot.
            - **Colors.** Displays the graph as a color bar using a gradient like the foreground and background colors.
        - **Color box.** Specifies the color of the graph for line and bar plots, and specifies a gradient for colors.

- **Gap fraction.** Which fraction of the sequences in the alignment that have gaps.
    - **Foreground color.** Colors the letter using a gradient, where the left side color is used if there are relatively few gaps, and the right side color is used if there are relatively many gaps.
    - **Background color.** Sets a background color of the residues using a gradient in the same way as described above.
    - **Graph.** Displays the gap fraction as a graph at the bottom of the alignment.
        - **Height.** Specifies the height of the graph.
        - **Type.** The type of the graph.
        - **Color box.** Specifies the color of the graph for line and bar plots, and specifies a gradient for colors.

- **Color different residues.** Indicates differences in aligned residues.
    - **Foreground color.** Colors the letter.
    - **Background color.** Sets a background color of the residues.

18.2.1 Sequence logo

Below the alignment there is an option of displaying a "sequence logo" (shown as default). The sequence logo displays the information content of all positions in the alignment as residues or nucleotides stacked on top of each other (see figure figure 18.8). The sequence logo provides a far more detailed view of the alignment than the conservation view (see section 18.2.2. Sequence logos can aid to identify protein binding sites on DNA sequences but can also aid to identify conserved residues in aligned domains of protein sequences and a wide range of other applications.

Each position of the alignment and consequently the sequence logo, show the sequence information in a computed score based on Shannon entropy [Schneider and Stephens, 1990]. The height of the individual letters represent the sequence information content in that particular position of the alignment.

A sequence logo is a much better visualization tool than a simple consensus sequence. An example hereof is for instance an alignment where in one position a particular residue is found in 70% of the sequences. If a consensus sequence would be defined it typically only displays the single residue with 70% coverage. In figure 18.8 and ungapped alignment of 11 *E. coli* start codons including flanking regions are shown. In this example, a consensus sequence would only display ATG as the start codon in position 1, but the looking at the sequence logo it is seen that a GTG is also allowed as a start codon.

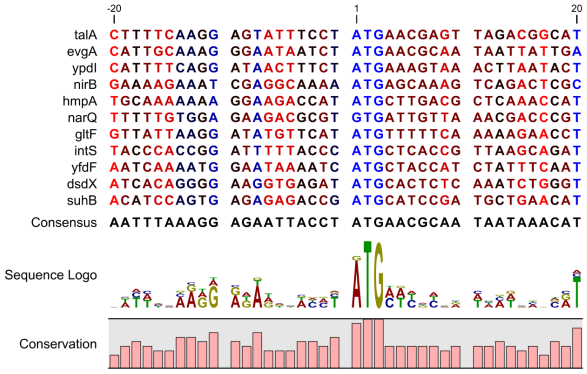


Figure 18.8: *Ungapped sequence alignment of eleven* E. coli *sequences defining a start codon. The start codons start at position 1. Below the alignment is shown the corresponding sequence logo. As seen, a GTG start codon and the usual ATG start codons are present in the alignment. This can also be visualized in the logo at position 1.*

Calculation of sequence logos

A comprehensive walk-through of the calculation of the information content in sequence logos is beyond the scope of this document but can be found in the original paper by [Schneider and Stephens, 1990]. Nevertheless, the conservation of every position is defined as R_{seq} which is the difference between the maximal entropy (S_{max}) and the observed entropy for the residue distribution (S_{obs}),

$$R_{seq} = S_{max} - S_{obs} = \log_2 N - \left(- \sum_{n=1}^{N} p_n \log_2 p_n \right)$$

p_n is the observed frequency of an amino acid residue or nucleotide of symbol n at a particular position and N is the number of distinct symbols for the sequence alphabet, either 20 for proteins or four for DNA/RNA. This means that the maximal sequence information content per position is $\log_2 4 = 2\ bits$ for DNA/RNA and $\log_2 20 \approx 4.32\ bits$ for proteins.

The original implementation by Schneider does not handle sequence gaps. We have slightly modified the algorithm so an estimated logo is presented in areas with sequence gaps.

If amino acid residues or nucleotides of one sequence are found in an area containing gaps, we have chosen to show the particular residue as the fraction of the sequences. Example; if one position in the alignment contain 9 gaps and only one alanine (A) the A represented in the logo has a hight of 0.1.

Other useful resources
The website of Tom Schneider
`http://www-lmmb.ncifcrf.gov/~toms/`

WebLogo
`http://weblogo.berkeley.edu/` [Crooks et al., 2004]

18.2.2 Conservation

The conservation view is very simplified view compared to the sequence logo view as described above. The bar (default view) show the conservation of all sequence positions. The height of

the bars in the view reflects how conserved that particular position is in the alignment. If one position is 100% conserved the bar will be shown in full height.

18.2.3 Gap fraction

The gap fraction view show if any gaps are present in the alignment. If a gap is present in the majority of sequences this will be represented in the view.

18.3 Edit alignments

18.3.1 Move residues and gaps

The placement of gaps in the alignment can be changed by modifying the parameters when creating the alignment (see section 18.1). However, gaps and residues can also be moved after the alignment is created:

select one or more gaps or residues in the alignment | **drag the selection to move**

This can be done both for single sequences, but also for multiple sequences by making a selection covering more than one sequence. When you have made the selection, the mouse pointer turns into a horizontal arrow indicating that the selection can be moved (see figure 18.9).

Notice! Residues can only be moved when they are next to a gap.

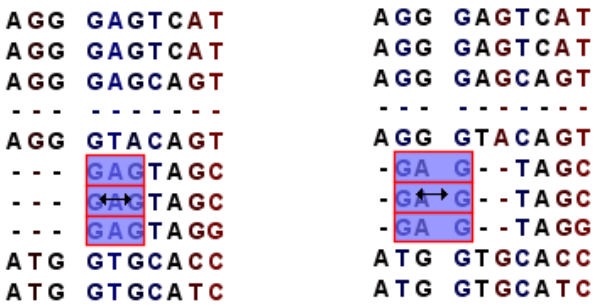


Figure 18.9: *Moving a part of an alignment. Notice the change of mouse pointer to a horizontal arrow.*

18.3.2 Insert gap columns

The placement of gaps in the alignment can be changed by modifying the parameters when creating the alignment. However, gaps can also be added manually after the alignment is created.

To insert extra gap columns (i.e. gaps in all the sequences):

select a part of the alignment | **right-click the selection** | **Add gap columns before/after**

If you have made a selection covering e.g. five residues, a gap of five will be inserted. In this way you can easily control the number of gaps to insert.

18.3.3 Delete residues and gaps

Residues or gaps can be deleted for individual sequences or for the whole alignment. For individual sequences:

> **select the part of the sequence you want to delete** | **right-click the selection** | **Edit selection** | **Delete the text in the dialog** | **Replace**

The selection shown in the dialog will be replaced by the text you enter. If you delete the text, the selection will be replaced by an empty text, i.e. deleted.

To delete entire columns:

> **select the part of the alignment you want to delete** | **right-click the selection** | **Delete columns**

The selection may cover one or more sequences, but the **Delete columns** function will always apply to the entire alignment.

18.3.4 Copy annotations to other sequences

Annotations on one sequence can be transferred to other sequences in the alignment:

> **right-click the annotation** | **Copy Annotation to other Sequences**

This will display a dialog listing all the sequences in the alignment. Next to each sequence is a checkbox which is used for selecting which sequences, the annotation should be copied to. Click **Copy** to copy the annotation.

18.3.5 Move sequences up and down

Sequences can be moved up and down in the alignment:

> **drag the label of the sequence up or down**

When you move the mouse pointer over the label, the pointer will turn into a vertical arrow indicating that the sequence can be moved.

The sequences can also be sorted automatically to let you save time moving the sequences around. To sort the sequences alphabetically:

> **Right-click the label of a sequence** | **Sort Sequences Alphabetically**

If you change the Sequence label (in the **Sequence Layout** view preferences), you will have to ask the program to sort the sequences again.

The sequences can also be sorted by similarity, grouping similar sequences together:

> **Right-click the label of a sequence** | **Sort Sequences by Similarity**

18.3.6 Delete and add sequences

Sequences can be removed from the alignment by right-clicking the label of a sequence:

> **right-click label** | **Delete Sequence**

This can be undone by clicking **Undo** ( ) in the Toolbar.

Extra sequences can be added to the alignment by creating a new alignment where you select the current alignment and the extra sequences (see section 18.1).

The same procedure can be used for joining two alignments.

18.3.7 Realign selection

If you have created an alignment, it is possible to realign a part of it, leaving the rest of the alignment unchanged:

select a part of the alignment to realign | **right-click the selection** | **Realign selection**

This will open **Step 2** in the "Create alignment" dialog, allowing you to set the parameters for the realignment (see section 18.1).

It is possible for an alignment to become shorter or longer as a result of the realignment of a region. This is because gaps may have to be inserted in, or deleted from, the sequences not selected for realignment. This will only occur for entire columns of gaps in these sequences, ensuring that their relative alignment is unchanged.

Realigning a selection is a very powerful tool for editing alignments in several situations:

- **Removing changes.** If you change the alignment in a specific region by hand, you may end up being unhappy with the result. In this case you may of course undo your edits, but another option is to select the region and realign it.
- **Adjusting the number of gaps.** If you have a region in an alignment which has too many gaps in your opinion, you can select the region and realign it. By choosing a relatively high gap cost you will be able to reduce the number of gaps.
- **Combine with fixpoints.** If you have an alignment where two residues are not aligned, but you know that they should have been. You can now set an alignment fixpoint on each of the two residues, select the region and realign it using the fixpoints. Now, the two residues are aligned with each other and everything in the selected region around them is adjusted to accommodate this change.

18.4 Join alignments

CLC Gene Workbench can join several alignments into one. This feature can for example be used to construct "supergenes" for phylogenetic inference by joining alignments of several disjoint genes into one spliced alignment. Note, that when alignments are joined, all their annotations are carried over to the new spliced alignment.

Alignments can be joined by:

select alignments to join | **Toolbox in the Menu Bar** | **Alignments and Trees( )** | **Join Alignments ( )**

or **select alignments to join** | **right-click either selected alignment** | **Toolbox** | **Alignments and Trees( )** | **Join Alignments ( )**

This opens the dialog shown in figure 18.10.

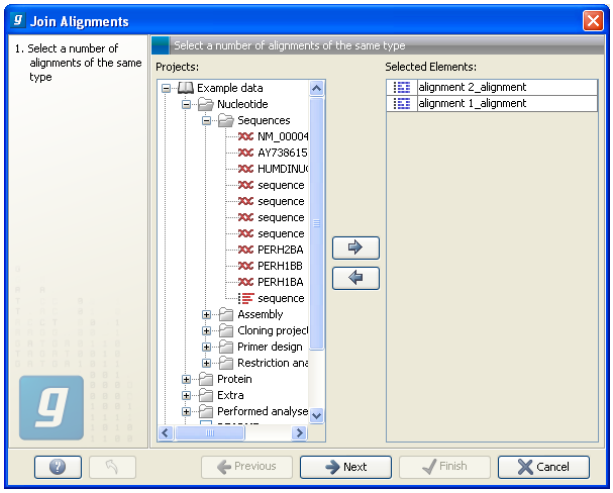


Figure 18.10: *Selecting two alignments to be joined.*

If you have selected some alignments before choosing the Toolbox action, they are now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove alignments from the **Project Tree**. Click **Next** opens the dialog shown in figure 18.11.

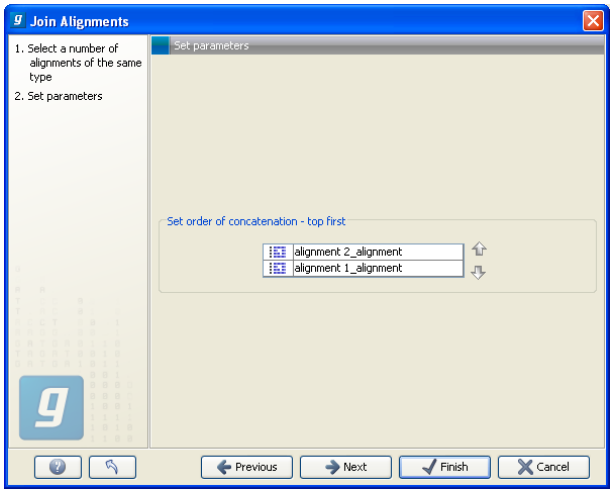


Figure 18.11: *Selecting order of concatenation.*

To adjust the order of concatenation, click the name of one of the alignments, and move it up or down using the arrow buttons.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. The result is seen in figure 18.12.

18.4.1 How alignments are joined

Alignments are joined by considering the sequence names in the individual alignments. If two sequences from different alignments have identical names, they are considered to have the same origin and are thus joined. Consider the joining of alignments A and B. If a sequence named "in-A-and-B" is found in both A and B, the spliced alignment will contain a sequence named "in-A-and-B" which represents the characters from A and B joined in direct extension of each other. If a sequence with the name "in-A-not-B" is found in A but not in B, the spliced alignment will contain a sequence named "in-A-not-B". The first part of this sequence will contain

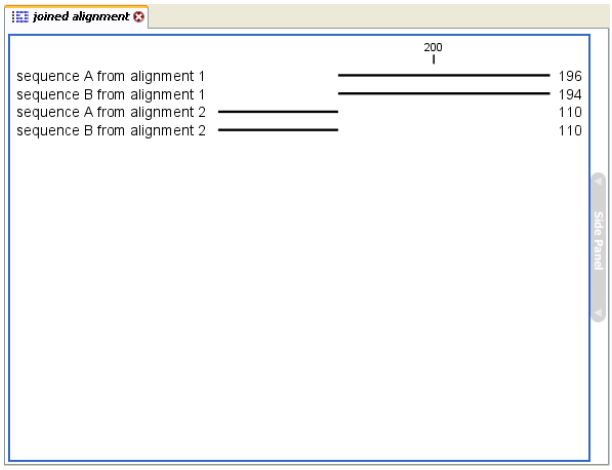


Figure 18.12: *The joining of the alignments result in one alignment containing rows of sequences corresponding to the number of uniquely named sequences in the joined alignments.*

the characters from A, but since no sequence information is available from B, a number of gap characters will be added to the end of the sequence corresponding to the number of residues in B. Note, that the function does not require that the individual alignments contain an equal number of sequences.

18.5 Bioinformatics explained: Multiple alignments

Multiple alignments are at the core of bioinformatical analysis. Often the first step in a chain of bioinformatical analyses is to construct a multiple alignment of a number of homologs DNA or protein sequences. However, despite their frequent use, the development of multiple alignment algorithms remains one of the algorithmically most challenging areas in bioinformatical research.

Constructing a multiple alignment corresponds to developing a hypothesis of how a number of sequences have evolved through the processes of character substitution, insertion and deletion. The input to multiple alignment algorithms is a number of homologous sequences i.e. sequences that share a common ancestor and most often also share molecular function. The generated alignment is a table (see figure 18.13) where each row corresponds to an input sequence and each column corresponds to a position in the alignment. An individual column in this table represents residues that have all diverged from a common ancestral residue. Gaps in the table (commonly represented by a '-') represent positions where residues have been inserted or deleted and thus do not have ancestral counterparts in all sequences.

18.5.1 Use of multiple alignments

Once a multiple alignment is constructed it can form the basis for a number of analyses:

- The phylogenetic relationship of the sequences can be investigated by tree-building methods based on the alignment.
- Annotation of functional domains, which may only be known for a subset of the sequences, can be transferred to aligned positions in other un-annotated sequences.
- Conserved regions in the alignment can be found which are prime candidates for holding functionally important sites.

- Comparative bioinformatical analysis can be performed to identify functionally important regions.

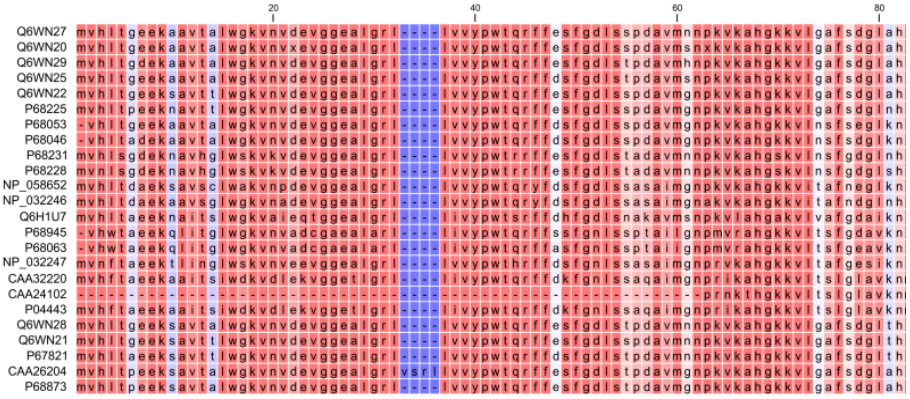


Figure 18.13: *The tabular format of a multiple alignment of 24 Hemoglobin protein sequences. Sequence names appear at the beginning of each row and the residue position is indicated by the numbers at the top of the alignment columns. The level of sequence conservation is shown on a color scale with blue residues being the least conserved and red residues being the most conserved.*

18.5.2 Constructing multiple alignments

Whereas the optimal solution to the pairwise alignment problem can be found in reasonable time, the problem of constructing a multiple alignment is much harder.

The first major challenge in the multiple alignment procedure is how to rank different alignments i.e. which *scoring function* to use. Since the sequences have a shared history they are correlated through their *phylogeny* and the scoring function should ideally take this into account. Doing so is, however, not straightforward as it increases the number of model parameters considerably. It is therefore commonplace to either ignore this complication and assume sequences to be unrelated, or to use heuristic corrections for shared ancestry.

The second challenge is to find the optimal alignment given a scoring function. For pairs of sequences this can be done by *dynamic programming* algorithms, but for more than three sequences this approach demands too much computer time and memory to be feasible.

A commonly used approach is therefore to do *progressive alignment* [Feng and Doolittle, 1987] where multiple alignments are built through the successive construction of pairwise alignments. These algorithms provide a good compromise between time spent and the quality of the resulting alignment

Presently, the most exciting development in multiple alignment methodology is the construction of *statistical alignment* algorithms [Hein, 2001], [Hein et al., 2000]. These algorithms employ a scoring function which incorporates the underlying phylogeny and use an explicit stochastic model of molecular evolution which makes it possible to compare different solutions in a statistically rigorous way. The optimization step, however, still relies on dynamic programming and practical use of these algorithms thus awaits further developments.

Creative Commons License
All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-

NoDerivs 2.5 License. You are free to to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and "CLC bio" has to be clearly labelled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.

See http://creativecommons.org/licenses/by-nc-nd/2.5/ for more about how you may use the contents.

Chapter 19

Phylogenetic trees

Contents

19.1 Inferring phylogenetic trees 256
19.1.1 Phylogenetic tree parameters 256
19.1.2 Tree View Preferences 258
19.2 Bioinformatics explained: phylogenetics 259
19.2.1 The phylogenetic tree 260
19.2.2 Modern usage of phylogenies 260
19.2.3 Reconstructing phylogenies from molecular data 261
19.2.4 Interpreting phylogenies 262

CLC Gene Workbench 2.0 offers different ways of inferring phylogenetic trees. The first part of this chapter will briefly explain the different ways of inferring trees in *CLC Gene Workbench 2.0*. The second part, "Bioinformatics explained", will give a more general introduction to the concept of phylogeny and the associated bioinformatics methods.

19.1 Inferring phylogenetic trees

For a given set of aligned sequences (see chapter 18) it is possible to infer their evolutionary relationships. In *CLC Gene Workbench 2.0* this is done by creating af phylogenetic tree:

Toolbox in the Menu Bar | **Alignments and Trees(** **)** | **Create Tree (** **)**

or **right-click alignment in Navigation Area** | **Toolbox** | **Alignments and Trees(** **)** | **Create Tree (****)**

This opens the dialog displayed in figure 19.1:

If an alignment was selected before choosing the Toolbox action, this alignment is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove elements from the **Navigation Area**. Click **Next** to adjust parameters.

19.1.1 Phylogenetic tree parameters

Figure 19.2 shows the parameters that can be set:

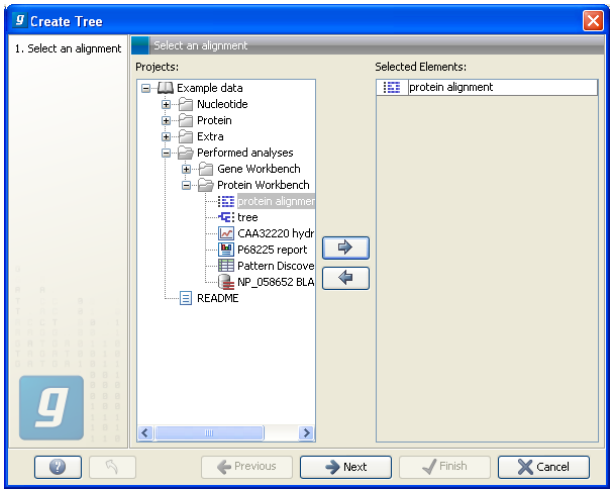


Figure 19.1: *Creating a Tree.*

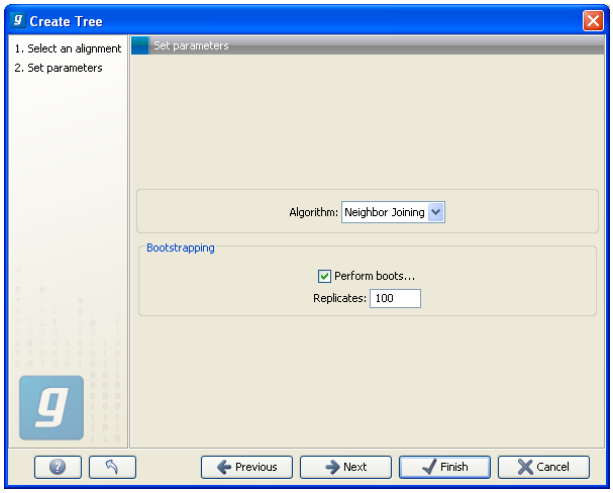


Figure 19.2: *Adjusting parameters.*

- Algorithms
  - The **UPGMA** method assumes that evolution has occured at a constant rate in the different lineages. This means that a root of the tree is also estimated.
  - The **neighbor joining** method builds a tree where the evolutionary rates are free to differ in different lineages. *CLC Gene Workbench 2.0* always draws trees with roots for practical reasons, but with the neighbor joining method, no particular biological hypothesis is postulated by the placement of the root. Figure 19.3 shows the difference between the two methods.
- To evaluate the reliability of the inferred trees, *CLC Gene Workbench 2.0* allows the option of doing a **bootstrap** analysis. A bootstrap value will be attached to each branch, and this value is a measure of the confidence in this branch. The number of replicates in the bootstrap analysis can be adjusted in the wizard. The default value is 100.

For a more detailed explanation, see "Bioinformatics explained" in section 19.2.

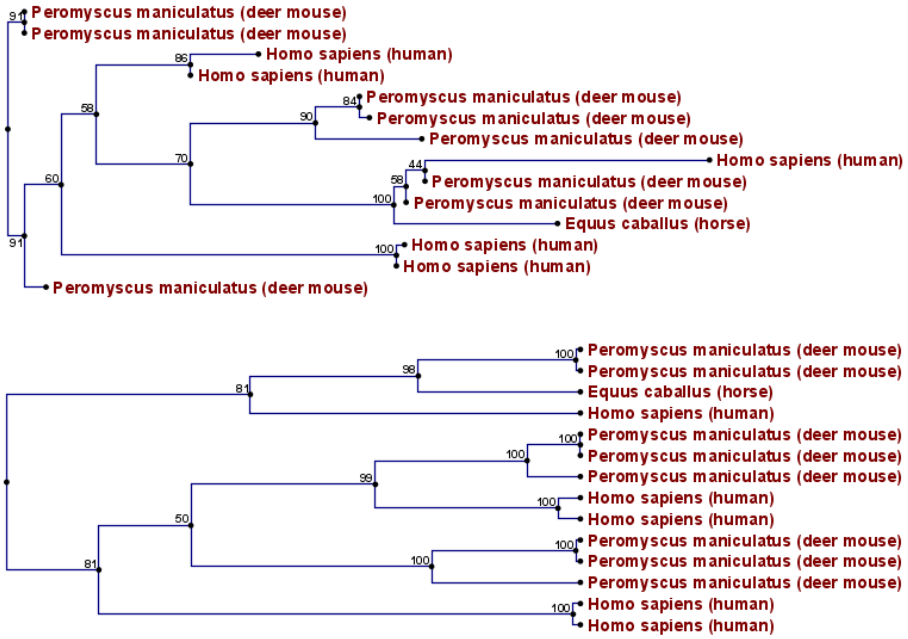


Figure 19.3: *Method choices for phylogenetic inference. The top shows a tree found by neighbor joining, while the bottom shows a tree found by UPGMA. The latter method assumes that the evolution occurs at a constant rate in different lineages.*

19.1.2 Tree View Preferences

The **Tree View** preferences are these:

- **Text format.** Changes the text format for all of the nodes the tree contains.
  - **Text size.** The size of the text representing the nodes can be modified in tiny, small, medium, large or huge.
  - **Font.** Sets the font of the text of all nodes
  - **Bold.** Sets the text bold if enabled.

- **Tree Layout.** Different layouts for the tree.
  - **Node symbol.** Changes the symbol of nodes into box, dot, circle or none if you don‘t want a node symbol.
  - **Layout.** Displays the tree layout as standard or topology.
  - **Show internal node labels**. This allows you to see labels for the internal nodes. Initially, there are no labels, but right-clicking a node allows you to type a label.
  - **Label color.** Changes the color of the labels on the tree nodes.
  - **Branch label color.** Modifies the color of the labels on the branches.
  - **Node color.** Sets the color of all nodes.
  - **Line color.** Alters the color of all lines in the tree.

- **Annotation Layout.** Specifies the annotation in the tree.

  - **Nodes.** Sets the annotation of all nodes either to name or to species.
  - **Branches.** Changes the annotation of the branches to bootstrap, length or none if you don‘t want annotation on branches.

Notice! Dragging in a tree will change it. You are therefore asked if you want to save this tree when the **Tree Viewer** is closed.

You may select part of a **Tree** by clicking on the nodes that you want to select.

Right-click a selected node opens a menu with the following options:

- Set root above node (defines the root of the tree to be just above the selected node).
- Set root at this node (defines the root of the tree to be at the selected node).
- Toggle collapse (collapses or expands the branches below the node).
- Change label (allows you to label or to change the existing label of a node).
- Change branch label (allows you to change the existing label of a branch).

You can also relocate leaves and branches in a tree or change the length.
Notice! To drag branches of a tree, you must first click the node one time, and then click the node again, and this time hold the mouse button.

In order to change the representation:

- Rearrange leaves and branches by
  Select a leaf or branch | **Move it up and down (Hint: The mouse turns into an arrow pointing up and down)**
- Change the length of a branch by
  Select a leaf or branch | **Press Ctrl** | **Move left and right (Hint: The mouse turns into an arrow pointing left and right)**

Alter the preferences in **Side Panel** for changing the presentation of the tree.
Notice! The preferences will not be saved. Viewing a tree in different viewers gives you the opportunity to change into different preferences in all of the viewers. For example if you select the **Annotation Layout** species for a node then you will only see the change in the specified view. If you now move leaves, the leaves in all views are moved. The options of the right-click pop up menu are changing the tree and therefore they change all views.

Notice! The **Set Root Above** and the **Set Root Here** functions change the tree, and therefore you may save it in order to be able to see it in this format later on.

19.2 Bioinformatics explained: phylogenetics

Phylogenetics describes the taxonomical classification of organisms based on their evolutionary history i.e. their *phylogeny*. Phylogenetics is therefore an integral part of the science of *systematics* that aims to establish the phylogeny of organisms based on their characteristics. Furthermore, phylogenetics is central to evolutionary biology as a whole as it is the condensation of the overall paradigm of how life arose and developed on earth.

19.2.1 The phylogenetic tree

The evolutionary hypothesis of a phylogeny can be graphically represented by a phylogenetic tree.

Figure 19.4 shows a proposed phylogeny for the great apes, *Hominidae*, taken in part from Purvis [Purvis, 1995]. The tree consists of a number of nodes (also termed vertices) and branches (also termed edges). These nodes can represent either an individual, a species, or a higher grouping and are thus broadly termed taxonomical units. In this case, the terminal nodes (also called leaves or tips of the tree) represent extant species of *Hominidae* and are the *operational taxonomical units* (OTUs). The internal nodes, which here represent extinct common ancestors of the great apes, are termed *hypothetical taxonomical units* since they are not directly observable.

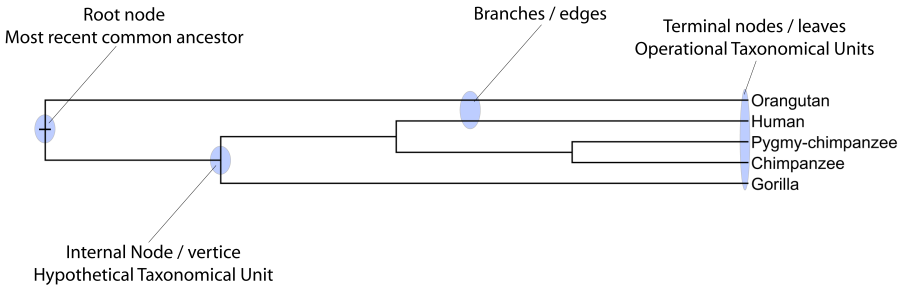


Figure 19.4: *A proposed phylogeny of the great apes (Hominidae). Different components of the tree are marked, see text for description.*

The ordering of the nodes determine the tree *topology* and describes how lineages have diverged over the course of evolution. The branches of the tree represent the amount of evolutionary divergence between two nodes in the tree and can be based on different measurements. A tree is completely specified by its topology and the set of all edge lengths.

The phylogenetic tree in figure 19.4 is rooted at the most recent common ancestor of all *Hominidae* species, and therefore represents a hypothesis of the direction of evolution e.g. that the common ancestor of gorilla, chimpanzee and man existed before the common ancestor of chimpanzee and man. If this information is absent trees can be drawn as unrooted.

19.2.2 Modern usage of phylogenies

Besides evolutionary biology and systematics the inference of phylogenies is central to other areas of research.

As more and more genetic diversity is being revealed through the completion of multiple genomes, an active area of research within bioinformatics is the development of comparative machine learning algorithms that can simultaneously process data from multiple species [Siepel and Haussler, 2004]. Through the comparative approach, valuable evolutionary information can be obtained about which amino acid substitutions are functionally tolerant to the organism and which are not. This information can be used to identify substitutions that affect protein function and stability, and is of major importance to the study of proteins [Knudsen and Miyamoto, 2001]. Knowledge of the underlying phylogeny is, however, paramount to comparative methods of inference as the phylogeny describes the underlying correlation from shared history that exists between data from different species.

In molecular epidemiology of infectious diseases, phylogenetic inference is also an important tool. The very fast substitution rate of microorganisms, especially the RNA viruses, means that these show substantial genetic divergence over the time-scale of months and years. Therefore, the phylogenetic relationship between the pathogens from individuals in an epidemic can be resolved and contribute valuable epidemiological information about transmission chains and epidemiologically significant events [Leitner and Albert, 1999], [Forsberg et al., 2001].

19.2.3 Reconstructing phylogenies from molecular data

Traditionally, phylogenies have been constructed from morphological data, but following the growth of genetic information it has become common practice to construct phylogenies based on molecular data, known as *molecular phylogeny*. The data is most commonly represented in the form of DNA or protein sequences, but can also be in the form of e.g. restriction fragment length polymorphism (RFLP).

Methods for constructing molecular phylogenies can be distance based or character based.

Distance based methods
Two common algorithms, both based on pairwise distances, are the UPGMA and the Neighbor Joining algorithms. Thus, the first step in these analyses is to compute a matrix of pairwise distances between OTUs from their sequence differences. To correct for multiple substitutions it is common to use distances corrected by a model of molecular evolution such as the Jukes-Cantor model [Jukes and Cantor, 1969].

UPGMA. A simple but popular clustering algorithm for distance data is Unweighted Pair Group Method using Arithmetic averages (UPGMA). [Michener and Sokal, 1957], [Sneath and Sokal, 1973]. This method works by initially having all sequences in separate clusters and continuously joining these. The tree is constructed by considering all initial clusters as leaf nodes in the tree, and each time two clusters are joined, a node is added to the tree as the parent of the two chosen nodes. The clusters to be joined are chosen as those with minimal pairwise distance. The branch lengths are set corresponding to the distance between clusters, which is calculated as the average distance between pairs of sequences in each cluster.

The algorithm assumes that the distance data has the so-called *molecular clock* property i.e. the divergence of sequences occur at the same constant rate at all parts of the tree. This means that the leaves of UPGMA trees all line up at the extant sequences and that a root is estimated as part of the procedure.

Neighbor Joining. The neighbor joining algorithm, [Saitou and Nei, 1987], on the other hand, builds a tree where the evolutionary rates are free to differ in different lineages, i.e., the tree does not have a particular root. Some programs always draw trees with roots for practical reasons, but for neighbor joining trees, no particular biological hypothesis is postulated by the placement of the root. The method works very much like UPGMA. The main difference is that instead of using pairwise distance, this method subtracts the distance to all other nodes from the pairwise distance. This is done to take care of situations where the two closest nodes are not neighbors in the ''real'' tree. The neighbor join algorithm is generally considered to be fairly good and is widely used. Algorithms that improves its cubic time performance exist. The improvement is only significant for quite large datasets.

Character based methods
Whereas the distance based methods compress all sequence information into a single number,

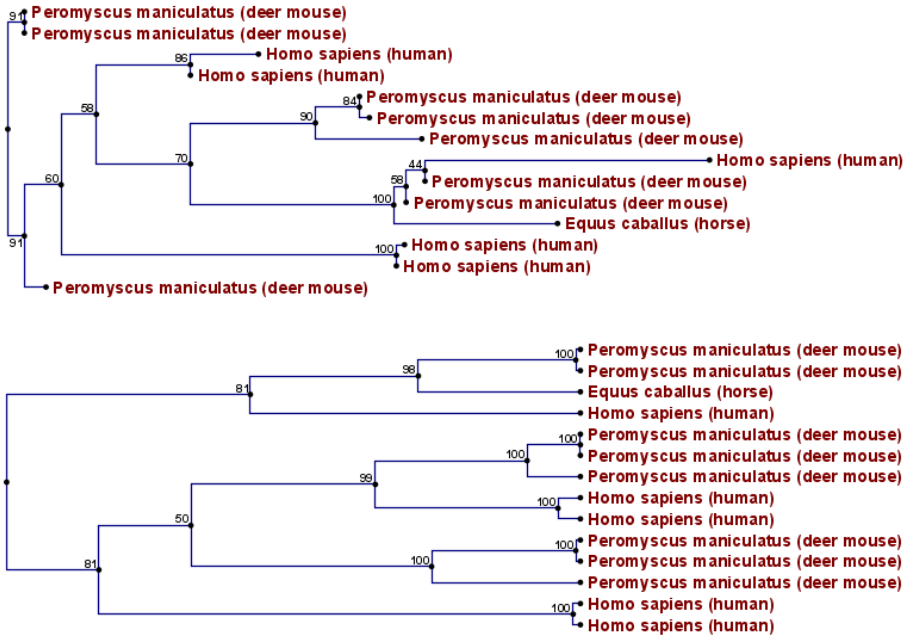


Figure 19.5: *Algorithm choices for phylogenetic inference. The top shows a tree found by the neighbor joining algorithm, while the bottom shows a tree found by the UPGMA algorithm. The latter algorithm assumes that the evolution occurs at a constant rate in different lineages.*

the character based methods attempt to infer the phylogeny based on all the individual characters (nucleotides or amino acids).

Parsimony. In parsimony based methods a number of sites are defined which are informative about the topology of the tree. Based on these, the best topology is found by minimizing the number of substitutions needed to explain the informative sites. Parsimony methods are not based on explicit evolutionary models.

Maximum Likelihood. Maximum likelihood and Bayesian methods (see below) are probabilistic methods of inference. Both have the pleasing properties of using explicit models of molecular evolution and allowing for rigorous statistical inference. However, both approaches are very computer intensive.

A stochastic model of molecular evolution is used to assign a probability (likelihood) to each phylogeny, given the sequence data of the OTUs. Maximum likelihood inference [Felsenstein, 1981] then consists of finding the tree which assign the highest probability to the data.

Bayesian inference. The objective of Bayesian phylogenetic inference is not to infer a single "correct" phylogeny, but rather to obtain the full posterior probability distribution of all possible phylogenies. This is obtained by combining the likelihood and the prior probability distribution of evolutionary parameters. The vast number of possible trees means that bayesian phylogenetics must be performed by approximative Monte Carlo based methods. [Larget and Simon, 1999], [Yang and Rannala, 1997].

19.2.4 Interpreting phylogenies

Bootstrap values

A popular way of evaluating the reliability of an inferred phylogenetic tree is bootstrap analysis.

The first step in a bootstrap analysis is to re-sample the alignment columns with replacement. I.e., in the re-sampled alignment, a given column in the original alignment may occur two or more times, while some columns may not be represented in the new alignment at all. The re-sampled alignment represents an estimate of how a different set of sequences from the same genes and the same species may have evolved on the same tree.

If a new tree reconstruction on the re-sampled alignment results in a tree similar to the original one, this increases the confidence in the original tree. If, on the other hand, the new tree looks very different, it means that the inferred tree is unreliable. By re-sampling a number of times it is possibly to put reliability weights on each internal branch of the inferred tree. If the data was bootstrapped a 100 times, a bootstrap score of 100 means that the corresponding branch occurs in all 100 trees made from re-sampled alignments. Thus, a high bootstrap score is a sign of greater reliability.

Other useful resources

The Tree of Life web-project
http://tolweb.org

Joseph Felsensteins list of phylogeny software
http://evolution.genetics.washington.edu/phylip/software.html

Creative Commons License
All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and "CLC bio" has to be clearly labelled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.

See http://creativecommons.org/licenses/by-nc-nd/2.5/ for more about how you may use the contents.

Part IV

Appendix

Appendix A

Comparison of workbenches

Below we list a number of functionalities that differ between CLC Workbenches:

- CLC Free Workbench (■)
- CLC Protein Workbench (■)
- CLC Gene Workbench (■)
- CLC Combined Workbench (■)

Batch processing	Free	Protein	Gene	Combined
Processing of multiple analyses in one single work-step		■	■	■

Database searches	Free	Protein	Gene	Combined
GenBank Entrez searches	■	■	■	■
UniProt searches (Swiss-Prot/TrEMBL)		■		■
Web-based sequence search using BLAST		■	■	■
PubMed searches		■	■	■
Web-based lookup of sequence data		■	■	■

General sequence analyses	Free	Protein	Gene	Combined
Linear sequence view	■	■	■	■
Circular sequence view	■	■	■	■
Text based sequence view	■	■	■	■
Editing sequences		■	■	■
Adding and editing sequence annotations		■	■	■
Sequence statistics	■	■	■	■
Shuffle sequence	■	■	■	■
Local complexity region analyses		■	■	■
Advanced protein statistics		■		■
Comprehensive protein characteristics report		■		■

For a more detailed comparison, we refer to http://www.clcbio.com.

Nucleotide analyses	Free	Protein	Gene	Combined
Basic gene finding	■	■	■	■
Reverse complement without loss of annotation	■	■	■	■
Restriction site analysis	■	■	■	■
Advanced interactive restriction site analysis			■	■
Translation of sequences from DNA to proteins	■	■	■	■
Interactive translations of sequences and alignments		■	■	■
G/C content analyses and graphs		■	■	■
Annotate with known SNP's in dbSNP database			■	■

Protein analyses	Free	Protein	Gene	Combined
3D molecule view		■		■
Hydrophobicity analyses		■	■	■
Antigenicity analysis		■		■
Protein charge analysis		■		■
Reverse translation from protein to DNA		■	■	■
Proteolytic cleavage detection		■		■
Prediction of signal peptides (SignalP)		■		■
Transmembrane helix prediction (TMHMM)		■		■
Secondary protein structure prediction		■		■
PFAM domain search		■		■

Sequence alignment	Free	Protein	Gene	Combined
Multiple sequence alignments (Two algorithms)	■	■	■	■
Advanced re-alignment and fix-point alignment options		■	■	■
Advanced alignment editing options		■	■	■
Consensus sequence determination and management	■	■	■	■
Conservation score along sequences	■	■	■	■
Sequence logo graphs along alignments		■	■	■
Gap fraction graphs		■	■	■

Dot plots	Free	Protein	Gene	Combined
Dot plot based analyses		■	■	■

Phylogenetic trees	Free	Protein	Gene	Combined
Neighbor-joining and UPGMA phylogenies	■	■	■	■

Pattern discovery	Free	Protein	Gene	Combined
Search for sequence match		■	■	■
Motif search		■	■	■
Pattern discovery		■	■	■

Primer design	Free	Protein	Gene	Combined
Advanced primer design tools			■	■
Detailed primer and probe parameters			■	■
Graphical display of primers			■	■
Generation of primer design output			■	■
Support for Standard PCR			■	■
Support for Nested PCR			■	■
Support for TaqMan PCR			■	■
Support for Sequencing primers			■	■
Match primer with sequence			■	■
Ordering of primers			■	■

Assembly of sequencing data	Free	Protein	Gene	Combined
Advanced contig assembly			■	■
Importing and viewing trace data			■	■
Trim sequences			■	■
Assemble without use of reference sequence			■	■
Assemble to reference sequence			■	■
Viewing and edit contigs			■	■

Molecular cloning	Free	Protein	Gene	Combined
Advanced molecular cloning			■	■
Graphical display of in silico cloning			■	■
Advanced sequence manipulation			■	■

Appendix B

BLAST databases

Several databases are available at NCBI, which can be selected to narrow down the possible BLAST hits.

B.1 Peptide sequence databases

- **nr.** Non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF, excluding those in env_nr.
- **refseq.** Protein sequences from NCBI Reference Sequence project http://www.ncbi.nlm.nih.gov/RefSeq/.
- **swissprot.** Last major release of the SWISS-PROT protein sequence database (no incremental updates).
- **pat.** Proteins from the Patent division of GenBank.
- **pdb.** Sequences derived from the 3-dimensional structure records from the Protein Data Bank http://www.rcsb.org/pdb/.
- **env_nr.** Non-redundant CDS translations from env_nt entries.
- **month.** All new or revised GenBank CDS translations + PDB + SwissProt + PIR + PRF released in the last 30 days..

B.2 Nucleotide sequence databases

- **nr.** All GenBank + EMBL + DDBJ + PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences). No longer "non-redundant" due to computational cost.
- **refseq_rna.** mRNA sequences from NCBI Reference Sequence Project.
- **refseq_genomic.** Genomic sequences from NCBI Reference Sequence Project.
- **est.** Database of GenBank + EMBL + DDBJ sequences from EST division.
- **est_human.** Human subset of est.

- **est_mouse.** Mouse subset of est.
- **est_others.** Subset of est other than human or mouse.
- **gss.** Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
- **htgs.** Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2. Finished, phase 3 HTG sequences are in nr.
- **pat.** Nucleotides from the Patent division of GenBank.
- **pdb.** Sequences derived from the 3-dimensional structure records from Protein Data Bank. They are NOT the coding sequences for the corresponding proteins found in the same PDB record.
- **month.** All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.
- **alu.** Select Alu repeats from REPBASE, suitable for masking Alu repeats from query sequences. See "Alu alert" by Claverie and Makalowski, Nature 371: 752 (1994).
- **dbsts.** Database of Sequence Tag Site entries from the STS division of GenBank + EMBL + DDBJ.
- **chromosome.** Complete genomes and complete chromosomes from the NCBI Reference Sequence project. It overlaps with refseq_genomic.
- **wgs.** Assemblies of Whole Genome Shotgun sequences.
- **env_nt.** Sequences from environmental samples, such as uncultured bacterial samples isolated from soil or marine samples. The largest single source is Sagarsso Sea project. This does overlap with nucleotide nr.

Appendix C

Formats for import and export

C.1 List of bioinformatic data formats

Below is a list of bioinformatic data formats, i.e. formats for importing and exporting sequences, alignments and trees.

File type	Suffix	File format used for
Phylip Alignment	.phy	alignments
GCG Alignment	.msf	alignments
Clustal Alignment	.aln	alignments
Newick	.nwk	trees
FASTA	.fsa/.fasta	sequences
GenBank	.gbk/.gb/.gp	sequences
GCG sequence	.gcg	sequences (only import)
PIR (NBRF)	.pir	sequences (only import)
Staden	.sdn	sequences (only import)
VectorNTI		sequences (only import)
DNAstrider	.str/.strider	sequences
Swiss-Prot	.swp	protein sequences
Lasergene sequence	.pro	protein sequence (only import)
Lasergene sequence	.seq	nucleotide sequence (only import)
Embl	.embl	nucleotide sequences
Nexus	.nxs/.nexus	sequences, trees, alignments, and sequence lists
CLC	.clc	sequences, trees, alignments, reports, etc.
Text	.txt	all data in a textual format
ABI		Trace files (only import)
AB1		Trace files (only import)
SCF2		Trace files (only import)
SCF3		Trace files (only import)
Phred		Trace files (only import)
mmCIF	.cif	structure (only import)
PDB	.pdb	structure (only import)
Preferences	.cpf	CLC workbench preferences

Notice that *CLC Gene Workbench* can import 'external' files, too. This means that *CLC Gene Workbench* can import all files and display them in the **Navigation Area**, while the above

mentioned formats are the types which can be read by *CLC Gene Workbench*.

C.2 List of graphics data formats

Below is a list of formats for exporting graphics. All data displayed in a graphical format can be exported using these formats. Data represented in lists and tables can only be exported in .pdf format (see section 6.3 for further details).

Format	Suffix	Type
Portable Network Graphics	.png	bitmap
JPEG	.jpg	bitmap
Tagged Image File	.tif	bitmap
PostScript	.ps	vector graphics
Encapsulated PostScript	.eps	vector graphics
Portable Document Format	.pdf	vector graphics
Scalable Vector Graphics	.svg	vector graphics

Bibliography

[Andrade et al., 1998] Andrade, M. A., O'Donoghue, S. I., and Rost, B. (1998). Adaptation of protein surfaces to subcellular location. *J Mol Biol*, 276(2):517–525.

[Bachmair et al., 1986] Bachmair, A., Finley, D., and Varshavsky, A. (1986). In vivo half-life of a protein is a function of its amino-terminal residue. *Science*, 234(4773):179–186.

[Bommarito et al., 2000] Bommarito, S., Peyret, N., and SantaLucia, J. (2000). Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic Acids Res*, 28(9):1929–1934.

[Cornette et al., 1987] Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A., and DeLisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J Mol Biol*, 195(3):659–685.

[Crooks et al., 2004] Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res*, 14(6):1188–1190.

[Dayhoff and Schwartz, 1978] Dayhoff, M. O. and Schwartz, R. M. (1978). *Atlas of Protein Sequence and Structure*, volume 3 of *5 suppl.*, chapter Atlas of Protein Sequence and Structure, pages 353–358. Nat. Biomed. Res. Found., Washington D.C.

[Eddy, 2004] Eddy, S. R. (2004). Where did the BLOSUM62 alignment score matrix come from? *Nat Biotechnol*, 22(8):1035–1036.

[Eisenberg et al., 1984] Eisenberg, D., Schwarz, E., Komaromy, M., and Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol*, 179(1):125–142.

[Engelman et al., 1986] Engelman, D. M., Steitz, T. A., and Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem*, 15:321–353.

[Felsenstein, 1981] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376.

[Feng and Doolittle, 1987] Feng, D. F. and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol*, 25(4):351–360.

[Forsberg et al., 2001] Forsberg, R., Oleksiewicz, M. B., Petersen, A. M., Hein, J., Bøtner, A., and Storgaard, T. (2001). A molecular clock dates the common ancestor of European-type porcine reproductive and respiratory syndrome virus at more than 10 years before the emergence of disease. *Virology*, 289(2):174–179.

[Gill and von Hippel, 1989] Gill, S. C. and von Hippel, P. H. (1989). Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem*, 182(2):319–326.

[Gonda et al., 1989] Gonda, D. K., Bachmair, A., Wünning, I., Tobias, J. W., Lane, W. S., and Varshavsky, A. (1989). Universality and structure of the N-end rule. *J Biol Chem*, 264(28):16700–16712.

[Hein, 2001] Hein, J. (2001). An algorithm for statistical alignment of sequences related by a binary tree. *Pacific symposium on biocomputing*, page 179.

[Hein et al., 2000] Hein, J., Wiuf, C., Knudsen, B., Møller, M. B., and Wibling, G. (2000). Statistical alignment: computational properties, homology testing and goodness-of-fit. *J Mol Biol*, 302(1):265–279.

[Henikoff and Henikoff, 1992] Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919.

[Hopp and Woods, 1983] Hopp, T. P. and Woods, K. R. (1983). A computer program for predicting protein antigenic determinants. *Mol Immunol*, 20(4):483–489.

[Ikai, 1980] Ikai, A. (1980). Thermostability and aliphatic index of globular proteins. *J Biochem (Tokyo)*, 88(6):1895–1898.

[Janin, 1979] Janin, J. (1979). Surface and inside volumes in globular proteins. *Nature*, 277(5696):491–492.

[Jukes and Cantor, 1969] Jukes, T. and Cantor, C. (1969). *Mammalian Protein Metabolism (ed. HN Munro)*, chapter Evolution of protein molecules, pages 21–32. New York: Academic Press.

[Knudsen and Miyamoto, 2001] Knudsen, B. and Miyamoto, M. M. (2001). A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc Natl Acad Sci U S A*, 98(25):14512–14517.

[Kyte and Doolittle, 1982] Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132.

[Larget and Simon, 1999] Larget, B. and Simon, D. (1999). Markov chain monte carlo algorithms for the bayesian analysis of phylogenetic trees. *Mol Biol Evol*, 16:750–759.

[Leitner and Albert, 1999] Leitner, T. and Albert, J. (1999). The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc Natl Acad Sci U S A*, 96(19):10752–10757.

[Maizel and Lenk, 1981] Maizel, J. V. and Lenk, R. P. (1981). Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc Natl Acad Sci U S A*, 78(12):7665–7669.

[McGinnis and Madden, 2004] McGinnis, S. and Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, 32(Web Server issue):W20–W25.

[Michener and Sokal, 1957] Michener, C. and Sokal, R. (1957). A quantitative approach to a problem in classification. *Evolution*, 11:130–162.

[Purvis, 1995] Purvis, A. (1995). A composite estimate of primate phylogeny. *Philos Trans R Soc Lond B Biol Sci*, 348(1326):405–421.

[Rose et al., 1985] Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., and Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science*, 229(4716):834–838.

[Saitou and Nei, 1987] Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425.

[SantaLucia, 1998] SantaLucia, J. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A*, 95(4):1460–1465.

[Schneider and Stephens, 1990] Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–6100.

[Siepel and Haussler, 2004] Siepel, A. and Haussler, D. (2004). Combining phylogenetic and hidden Markov models in biosequence analysis. *J Comput Biol*, 11(2-3):413–428.

[Sneath and Sokal, 1973] Sneath, P. and Sokal, R. (1973). *Numerical Taxonomy*. Freeman, San Francisco.

[Tobias et al., 1991] Tobias, J. W., Shrader, T. E., Rocap, G., and Varshavsky, A. (1991). The N-end rule in bacteria. *Science*, 254(5036):1374–1377.

[Wootton and Federhen, 1993] Wootton, J. C. and Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers iin Chemistry*, 17:149–163.

[Yang and Rannala, 1997] Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol Biol Evol*, 14(7):717–724.

Part V

Index

Index

AB1, file format, 29, 86, 270
ABI, file format, 29, 86, 270
About CLC Workbenches, 18
Accession number, display, 61
Activate license
    commercial, 17
    demo, 16
Add
    annotations, 124, 265
    enzymes cutting selection, 120, 133, 224
    sequences to alignment, 251
    sequences to contig, 216
Advanced preferences, 78
Algorithm
    alignment, 240
    neighbor joining, 261
    UPGMA, 261
Align
    alignments, 243
    protein sequences, tutorial, 32
    sequences, 265
Alignment Primers
    Degenerate primers, 201, 202
    PCR primers, 200
    Primers with mismatches, 201, 202
    Primers with perfect match, 201, 202
    TaqMan Probes, 200
Alignment-based primer design, 200
Alignments, 240, 265
    add sequences to, 251
    create, 241
    design primers for, 200
    edit, 249
    fast algorithm, 242
    join, 251
    multiple, Bioinformatics explained, 253
    remove sequences from, 250
    view, 246
Aliphatic index, 153
.aln, file format, 86
Ambiguities, reverse translation, 182
Amino acid composition, 155
Analyze primer properties, 204
Annotate with SNP's, 265
Annotation
    add, 124
    copy to other sequences, 250
    edit, 124
    in alignments, 250
    layout, 119
    map, 128
    overview, 128
    trim, 210
    types, 120
Antigenicity, 265
Append wildcard, search, 102
Arrange
    layout of sequence, 29
    views in View Area, 67
Assemble
    sequences, 212
    to existing contig, 216
    to reference sequence, 213
Assembly, 265
    tutorial, 45
    variance table, 219
Atomic composition, 154
Automatic parsing, 87

Back up, 90
Base pairs
    required for a match, 206
    required for mispriming, 193
Basic concepts of use, 20
Batch processing, 97, 265
    log of, 98
Bibliography, 272
Binding site for primer, 205
Bioinformatic data
    export, 88
    formats, 85, 270
BLAST, 265

against local Database, 113
against NCBI, 107
create database from file system, 114
create database from Navigation Area, 114
create local database, 114
graphics output, 111
list of databases, 268
parameters, 109
search, 107
table output, 112
tutorial, 38
BLAST DNA sequence
BLASTn, 108
BLASTx, 108
tBLASTx, 108
BLAST Protein sequence
BLASTp, 108
tBLASTn, 108
BLOSUM, scoring matrices, 142
Bootstrap values, 262
Bug reporting, 19

C/G content, 122
CDS, translate to protein, 124
Cheap end gaps, 242
.cif, file format, 86
Circular molecules, 229
Circular view of sequence, 132, 265
.clc, file format, 86, 89
CLC Standard Settings, 78, 79
CLC Workbenches, 18
CLC, file format, 29, 86, 270
Cloning, 221, 265
circular view, 229
insert fragment, 228
navigation, 224
restriction enzymes, 228
view preferences, 223
Close View, 65
Clustal, file format, 29, 86, 270
Coding sequence, translate to protein, 124
Codon
frequency tables, reverse translation, 181
usage, 182
Color residues, 247
Compare workbenches, 265
Complexity plot, 147
Configure network, 22
Conflicts, overview in assembly, 219
Consensus sequence, 246, 265
open, 246
Conservation, 246
graphs, 265
Contact information, 11
Contig, 265
ambiguities, 219
create, 212
view and edit, 216
Convert old data, 87
Copy, 93
annotations in alignments, 250
elements in Navigation Area, 60
into sequence, 124
search results, GenBank, 104
sequence, 128, 130
sequence selection, 165
text selection, 128
.cpf, file format, 78
Create
a project, tutorial, 26
alignment, 241
dot plots, 137
enzyme list, 234
local BLAST database, 114
new folder, 59
new project, 59
workspace, 73

Data formats
bioinformatic, 270
graphics, 271
Data structure, 58
Database
GenBank, 101
local, 58
nucleotide, 268
peptide, 268
Delete
element, 62
residues and gaps in alignment, 250
workspace, 73
Demo license, 15
Dipeptide distribution, 155
DNA translation, 166
DNAstrider, file format, 29, 86, 270
Dot plots, 265
Bioinformatics explained, 139
create, 137

print, 138
Double stranded DNA, 118
Download and open
search results, GenBank, 104
Download and save
search results, GenBank, 104
Download of *CLC Gene Workbench*, 11
Drag and drop, 48
Navigation Area, 60
search results, GenBank, 104

Edit
alignments, 249, 265
annotations, 124, 265
enzymes, 120, 133, 224
sequence, 124
sequences, 265
Element, 58
delete, 62
rename, 62
.embl, file format, 86
Embl, file format, 29, 86, 270
Encapsulated PostScript, export, 92
End gap cost, 242
End gap costs
cheap end caps, 242
free end gaps, 242
Enzyme list
create, 234
modify, 235
.eps-format, export, 92
Error reports, 19
Evolutionary relationship, 256
Example data, import, 22
Expect, BLAST search, 111
Export
bioinformatic data, 88
dependent objects, 89
folder, 88
graphics, 91
history, 89
list of formats, 270
multiple files, 88
preferences, 78
project, 88
External files, import and export, 90
Extinction coefficient, 153
Extract sequences, 132

FASTA, file format, 29, 86, 270
Feature request, 19
Feature table, 155
Features, *see* Annotation
File system, local BLAST database, 114
Find open reading frames, 168
Fit Width, 71
Fixpoints, for alignments, 244
Floating Side Panel, 79
Format, of the manual, 24
FormatDB, 114
Fragments, separate on gel, 237
Free end gaps, 242
.fsa, file format, 86

G/C content, 122, 265
G/C restrictions
3' end of primer, 188
5' end of primer, 189
End length, 188
Max G/C, 188
Gap
delete, 250
extension cost, 242
fraction, 247, 265
insert, 249
open cost, 242
.gbk, file format, 86
GC content, 187
GCG Alignment, file format, 29, 86, 270
GCG Sequence, file format, 29, 86, 270
Gel electrophoresis, 235
marker, 238
view, 237
view preferences, 237
when finding restriction sites, 232
GenBank
file format, 29, 86, 270
search, 101, 265
search sequence in, 105
tutorial, 31
Gene finding, 168
General preferences, 77
General Sequence Analyses, 136
Genetic code, reverse translation, 181
Getting started, 20
Google sequence, 105
Graphics
data formats, 271

export, 91

Half-life, 153
Handling of results, 97
Help, 20
Hide/show Toolbox, 72
History, 95
export, 89
preserve when exporting, 96
source elements, 96
Hydrophobicity, 174, 265
Bioinformatics explained, 177
Cornette, 178
Eisenberg, 178
Engelman (GES), 178
Hopp-Woods, 178
Janin, 178
Kyte-Doolittle, 177
Rose, 178

Import
bioinformatic data, 86
data from older versions, 87
existing data, 27
external files, 90
FASTA-data, 27
list of formats, 270
preferences, 78
Vector NTI data, 87
Infer Phylogenetic Tree, 256
Insert
gaps, 249
restriction site, 120, 133, 224
Installation, 11
Isoelectric point, 153

Join
alignments, 251
sequences, 156
.jpg-format, export, 92

Lasergene sequence
protein file format, 29, 86, 270
sequence file format, 29, 86, 270
License, 15
Linux
installation, 13
installation with RPM-package, 14
List of sequences, 130
Load enzymes, 120, 133, 224
Local BLAST Database, 114
Local complexity plot, 147, 265
Local Database, BLAST, 113
Locale setting, 77
Location
of selection on sequence, 71
Side Panel, 77
Log of batch processing, 98
Logo, sequence, 246, 265

Mac OS X installation, 13
Manipulate sequences, 265
Manual format, 24
Marker, in gel view, 238
Max Sequence length for BLAST, 107
Maximize size of view, 67
Maximum memory, adjusting, 22
Melting temperature, 188
Cation concentration, 188
Inner, 188
Primer concentration, 188
Memory, adjust maximum amount, 22
Menu Bar, illustration, 58
mmCIF, file format, 29, 86, 270
Mode toolbar, 69
Modify enzyme list, 235
Molecular weight, 152
Motif search, 157, 265
Mouse modes, 69
Move
content of a view, 71
elements in Navigation Area, 60
sequences in alignment, 250
.msf, file format, 86
Multiple alignments, 253, 265
Multiselecting, 60

Navigation Area, 58
create local BLAST database, 114
illustration, 58
NCBI, 101
search sequence in, 105
search, tutorial, 31
Negatively charged residues, 154
Neighbor Joining algorithm, 261
Neighbor-joining, 265
Nested PCR primers, 265
Network configuration, 22
New

feature request, 19
folder, 26, 59
project, 26, 59
sequence, 129
Newick, file format, 29, 86, 270
.nexus, file format, 86
Nexus, file format, 29, 86, 270
Non-standard residues, 121
nr, BLAST databases, 109
Nucleotide
info, 121
sequence databases, 268
Numbers on sequence, 118
.nwk, file format, 86
.nxs, file format, 86

Old data, import, 87
Online check, of demo license key, 15
Open
consensus sequence, 246
files, 20
Open reading frame determination, 168
Open-ended sequence, 168
Order primers, 207, 265
ORF, 168
Origins from, 96

Page setup, 83
PAM, scoring matrices, 142
Parameters
search, 102
Parsing, automatic, 87
Paste/copy, 93
Pattern Discovery, 160
Pattern discovery, 265
Pattern Search, 157
PCR primers, 265
.pdb, file format, 86
.seq, file format, 86
PDB, file format, 29, 86, 270
.pdf-format, export, 92
Peptide sequence databases, 268
Personal information, 19
Pfam domain search, 265
Phred, file format, 29, 86, 270
.phy, file format, 86
Phylip, file format, 29, 86, 270
Phylogenetic tree, 256, 265
tutorial, 34
Phylogenetics, Bioinformatics explained, 259
.pir, file format, 86
PIR (NBRF), file format, 29, 86, 270
Plot
dot plot, 137
local complexity, 147
.png-format, export, 92
Polarity colors, 121
Positively charged residues, 155
PostScript, export, 92
Preferences, 76
advanced, 78
export, 78
General, 77
import, 78
style sheet, 78
toolbar, 77
View, 77
view, 68
Primer, 205
analyze, 204
based on alignments, 200
design, 265
display graphically, 189
length, 187
mode, 189
nested PCR, 189
order, 207
sequencing, 189
standard, 189
TaqMan, 189
tutorial, 41
Print, 82
dot plots, 138
preview, 83
visible area, 82
whole view, 82
.pro, file format, 86
Problems when starting up, 19
Processes, 72
Project, create new, 26
Protein
charge, 171, 265
hydrophobicity, 177
Isoelectric point, 153
report, 265
statistics, 152
translation, 179

Proteolytic cleavage, 265
Proxy server, 22
Proxy settings
and license activation, 15
.ps-format, export, 92
PubMed references, search, 105
PubMed references,search, 265

Quality of trace, 210
Quick start, 21

Rasmol colors, 121
Reading frame, 168
Realign alignment, 265
Rebase, restriction enzyme database, 234
Recycle Bin, 62
Redo alignment, 243
Redo/Undo, 66
Reference sequence, 265
References, 272
Region
syntax, 125
types, 126
Remove
annotations, 126
sequences from alignment, 250
terminated processes, 72
Rename element, 62
Replace file, 91
Report program errors, 19
Report, protein, 265
Request new feature, 19
Reset license, 17, 18
Residue coloring, 121
Restore
deleted elements, 62
size of view, 68
Restriction enzymes, 230
separate on gel, 237
Restriction sites, 230, 265
enzyme database Rebase, 234
circular DNA, 133
on sequence, 120
parameters, 230
tutorial, 35
Results handling, 97
Reverse complement, 165, 265
Reverse translation, 179, 265
Bioinformatics explained, 181
RNA translation, 166

Safe mode, 19
Save
changes in a view, 66
search, 32
sequence, 32
style sheet, 78
view preferences, 78
workspace, 73
SCF2, file format, 29, 86, 270
SCF3, file format, 29, 86, 270
Score, BLAST search, 111
Scoring matrices
Bioinformatics explained, 142
BLOSUM, 142
PAM, 142
Search
BLAST, 107
GenBank, 101
handle results from GenBank, 103
hits, number of, 77
in a sequence, 122
in annotations, 122
Local BLAST, 113
options, GenBank, 101
parameters, 102
patterns, 157, 160
PubMed references, 105
sequence in UniProt, 106
sequence on Google, 105
sequence on NCBI, 105
sequence on web, 104
Secondary structure prediction, 265
Secondary structure, for primers, 188
Select
exact positions, 122
in sequence, 123
parts of a sequence, 123
workspace, 73
Selection mode in the toolbar, 71
Selection, location on sequence, 71
Self annealing, 188
Self end annealing, 188
Separate sequences on gel, 236
using restriction enzymes, 237
Sequence
alignment, 240
analysis, 136

display different information, 61
extract from sequence list, 132
information, 127
information, tutorial, 37
join, 156
layout, 118
lists, 130
logo, 265
new, 129
region types, 126
search, 122
select, 123
shuffle, 146
statistics, 149
view, 117
view as text, 128
view circular, 132
view format, 61
web info, 104
Sequence logo, 246, 247
Sequencing data, 265
Sequencing primers, 265
Shortcuts, 74
Show/hide Toolbox, 72
Shuffle sequence, 146, 265
Side Panel, location of, 77
Signal peptide, 265
SNP, 167
annotation, 167, 265
Sort
sequences, 132
sequences alphabetically, 250
sequences by similarity, 250
Source element, 96
Species, display sequence species, 61
Staden, file format, 29, 86, 270
Standard layout, trees, 259
Standard Settings, CLC, 79
Start Codon, 168
Start-up problems, 19
Statistics
about sequence, 265
protein, 152
sequence, 149
Status Bar, 72, 73
illustration, 58
.str, file format, 86
Style sheet, preferences, 78
Support mail, 11
.svg-format, export, 92
Swiss-Prot, file format, 29, 86, 270
Swiss-Prot/TrEMBL, 265
.swp, file format, 86
System requirements, 14

Tabs, use of, 64
TaqMan primers, 265
tBLASTn, 108
tBLASTx, 108
Terminated processes, 72
Text format, 123
user manual, 24
view sequence, 128
Text, file format, 29, 86, 270
.tif-format, export, 92
Tips and tricks, tutorial, 47
Toolbar
illustration, 58
preferences, 77
Toolbox, 72
illustration, 58
show/hide, 72
Topology layout, trees, 259
Trace data, 208, 265
quality, 210
Translate
a selection, 121
along DNA sequence, 121
annotation to protein, 124
CDS, 167
coding regions, 167
DNA to RNA, 163
nucleotide sequence, 166
ORF, 167
protein, 179
RNA to DNA, 164
to DNA, 265
to protein, 166, 265
Translation
of a selection, 121
show together with DNA sequence, 121
tables, 166
Transmembrane helix prediction, 265
Trim, 209, 265
.txt, file format, 86

Undo limit, 77

Undo/Redo, 66
UniProt
    search, 265
    search sequence in, 106
UniVec, trimming, 210
UPGMA algorithm, 261, 265
Upgrade license, 18
Urls, Navigation Area, 90
User defined view settings, 78
User interface, 58

Variance table, assembly, 219
Vector
    see cloning, 221
Vector contamination, find automatically, 210
Vector design, 221
Vector graphics, export, 92
VectorNTI
    file format, 29, 86, 270
    import data from, 87
View, 64
    alignment, 246
    dot plots, 138
    preferences, 68
    save changes, 66
    sequence, 117
    sequence as text, 128
View Area, 64
    illustration, 58
View preferences, 77
    show automatically, 77
    style sheet, 78
View settings
    user defined, 78

Wildcard, append to search, 102
Windows installation, 12
Workspace, 73
    create, 73
    delete, 73
    save, 73
    select, 73
Wrap sequences, 118

Zoom, 69
    tutorial, 29
Zoom In, 69
Zoom Out, 71
Zoom to 100% , 71